es

European Seminar

University of Cyprus

BIOSTAT2006

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE

STATISTICAL METHODS FOR BIOMEDICAL AND TECHNICAL SYSTEMS

Limassol, Cyprus

Edited By

Filia Vonta

Nicosia 2006

Proceedings of the International Conference
"Biostat 2006 – Statistical Methods for
Biomedical and Technical Systems"
May 2006

ISBN: 9963-644-53-8

Printing and Binding
Kantzilaris Bookshops Ltd
P.O.Box 20774 Nicosia 1663 Cyprus

Preface

This volume consists of accepted papers that will be presented at BIOSTAT2006, an international conference organized by the University of Cyprus and the European Seminar - Seminaire Europeen "Méthodes Mathématiques pour l' Analyse de Survie, Fiabilité et Qualite de Vie". In fact, this conference is a part of a series of conferences, workshops and seminars organized or co-organized by the European Seminar over the years. Our objective in taking part in this organization was to bring together scientists from all over the world that work in Statistics in general and advance the knowledge in fields related to biomedical and technical systems. During the conference a special workshop will be presented by Prof. Mei-Ling Ting Lee on the topic of Microarray Data.

The papers included in this volume represent a cross-section of current concerns and research interests in survival analysis, reliability, medical statistics, biology and epidemiology. The papers are presented in alphabetic order within each category (keynote, invited or contributed) using the name of the first author.

The editor would like to thank all the authors whose contribution made the publication of this book possible. The members of the scientific and organizing committees helped enormously in turning this effort into a very successful event. I also like to thank the colleagues that organized invited sessions or invited talks. The materialization of this conference would not have been possible without the constant support and guidance of Prof. Misha Nikulin and the organizational skills of Prof. Alex Karagrigoriou. Special thanks are due to the University of Cyprus for embracing warmly the organization of this event. Last but not least I like to thank all our sponsors that accepted our invitation to support the conference without any hesitation.

Filia Vonta

Limassol, May 2006

Conference Chairs

M. Nikulin, France

F. Vonta, Cyprus

Scientific Committee

N. Balakrishnan, Canada

C. Huber-Carol, France

M.-L. T. Lee, USA

N. Limnios, France

M. Nikulin, France

F. Vonta, Cyprus

Organizing Committee

A. Constantinou, Cyprus

A. Karagrigoriou, Cyprus

M. Nikulin, France

F. Vonta, Cyprus

Contents

Part I: Keynote Papers

Part II: Invited Papers

Part I

Keynote Papers

1

Hypotheses Testing: Poisson Versus Self-Exciting and Self-Correcting

S. Dachian and Yu. A. Kutoyants

Laboratoire de Mathématiques, Université Blaise Pascal, Aubière, France
Laboratoire "Statistique & Processus", Université du Maine, Le Mans, France

Abstract: We consider the problems of hypotheses testing with the basic simple hypothesis: observed sequence of points corresponds to a stationary Poisson process with known intensity. The alternatives are stationary self-exciting and self-correcting point processes. In the case of self-exciting alternatives we propose asymptotically uniformly most powerful tests in parametric and nonparametric statements and for the self-correcting alternative we compare the score-function test, likelihood ratio test and Wald's test with the Neyman-Pearson test. The results of numerical simulations of the tests are presented.

Keywords and phrases: Poisson process, self-exciting process, self-correcting process, hypotheses testing, asymptotically uniformly most powerful test

1.1 Introduction

Let $\{t_1, t_2, \ldots\}$ be a sequence of events of a stationary point process $X = \{X_t,\ t \geq 0\}$ (X_t is a counting process). The simplest stationary point process is, of course, Poisson process with a constant intensity $S > 0$, i.e., the increments of X on disjoint intervals are independent and distributed according to Poisson law

$$\mathbf{P}\left\{X_t - X_s = k\right\} = \frac{S^k\,(t-s)^k}{k!}\,\mathrm{e}^{-S(t-s)}, \quad 0 \leq s < t, \quad k = 0, 1, \ldots$$

Note that the statistical inference for Poisson processes is relatively simple (see Kutoyants (1998)). Therefore if we have a stationary sequence of events it is interesting to check first of all if this model (Poisson process) corresponds well to the observations (see Cox and Lewis (1966)). Note that any inhomogeneous Poisson process with known intensity can be transformed into homogeneous

Poisson process by time-changing. As alternatives we consider two types of stationary point processes: *self-exciting process* introduced by Hawkes (1972) and *self-correcting process* introduced by Isham and Westcott (1979). The advantage of these models is in the possibility to apply the likelihood ratio analysis, because the intensity functions depend of the observations. The detailed discussion of these processes can be found in Daley and Vere-Jones (2003).

1.2 Self-Exciting Alternatives

The self-exciting point process is defined by intensity function of the following form

$$S\left(t,X\right)=S_{*}+\int_{-\infty}^{t}g\left(t-s\right)\ \mathrm{d}X_{s},\qquad\int_{0}^{\infty}g\left(t\right)\ \mathrm{d}t<1$$

where $S_{*}>0$ and the function $g\left(\cdot\right)\geq0$. For example, $g\left(t\right)=\alpha e^{-\gamma t}$ with $\alpha/\gamma<1$. The basic hypothesis corresponds to $g\left(t\right)\equiv0$. We consider two situations. The first one is parametric, when $g\left(t\right)=\vartheta h\left(t\right)$, where $h\left(t\right)$ is known function and the second is nonparametric, when $g\left(t\right)$ is unknown function. In both cases we reparametrize the models to have contigual (asymptotically non degenerate) alternatives.

Note that self-exciting processes cover a large class of stationary point processes with rational spectral density (see, for example, Pham (1981)).

1.2.1 One-sided parametric alternative

We assume that the observed process is either Poisson with constant known intensity S_{*} or it is self-exciting processes with intensity function $\left(\vartheta=u/\sqrt{T}\right)$

$$S\left(\vartheta,t,\omega\right)=S_{*}+\frac{u}{\sqrt{T}}\int_{-\infty}^{t}h\left(t-s\right)\ \mathrm{d}X_{s},$$

i.e., we have to test the following two hypotheses

$$\begin{aligned}\mathcal{H}_{0}\ &:\qquad u=0,\\ \mathcal{H}_{1}\ &:\qquad u>0.\end{aligned}$$

Let us denote

$$\Delta_{T}\left(X^{T}\right)=\frac{1}{S_{*}\sqrt{T}}\int_{0}^{T}\int_{0}^{t}h\left(t-s\right)\ \mathrm{d}X_{s}\ \left[\mathrm{d}X_{t}-S_{*}\ \mathrm{d}t\right]$$

and put $c_{\varepsilon}=z_{\varepsilon}\ \sqrt{\mathrm{I}_{h}^{*}}$. Here

$$\mathrm{I}_{h}^{*}=\int_{0}^{\infty}h\left(t\right)^{2}\ \mathrm{d}t+S_{*}\ \left(\int_{0}^{\infty}h\left(t\right)\ \mathrm{d}t\right)^{2}.$$

A test $\phi_T^*(\cdot)$ is called asymptotically uniformly most powerful (AUMP) in the class $\mathcal{K}_\varepsilon$ of tests of asymptotic level $1-\varepsilon$ if for any other test $\phi_T(\cdot) \in \mathcal{K}_\varepsilon$ and any constant $K > 0$ we have

$$\lim_{T\to\infty} \inf_{0<u\le K} \left[\beta_T^*(u, \phi_T) - \beta_T(u, \phi_T)\right] \ge 0.$$

Theorem 1.2.1 *Let $h(\cdot) \in \mathcal{L}^1(R_+) \cap \mathcal{L}^2(R_+)$ then the test*

$$\hat{\phi}_T\left(X^T\right) = \chi_{\{\Delta_T(X^T) > c_\varepsilon\}}$$

is asymptotically uniformly most powerful in the class $\mathcal{K}_\varepsilon$ and for any $u > 0$ the power function

$$\beta_T\left(u, \hat{\phi}_T\right) \longrightarrow \hat{\beta}(u) = \mathbf{P}\left\{\zeta > z_\varepsilon - u\sqrt{\mathrm{I}_h^*}\right\},$$

where $\zeta \sim \mathcal{N}(0,1)$.

1.2.2 One-sided nonparametric alternative

We suppose that under hypothesis $\mathcal{H}_0$ the observed point process X^T is standard Poisson with known intensity $S_* > 0$ and under alternative $\mathcal{H}_1$ the intensity function is

$$S(t, X) = S_* + \frac{1}{\sqrt{T}} \int_{-\infty}^{t} u(t-s)\ \mathrm{d}X_s, \qquad 0 \le t \le T$$

where the function $u(\cdot) \ge 0$ is from the set

$$\mathcal{U}_\varrho = \left\{u(\cdot): \quad \int_0^\infty u(t)\ \mathrm{d}t = \varrho\right\}$$

with some $\rho > 0$.

A test $\phi_T^*(\cdot)$ is called locally asymptotically uniformly most powerful in the class $\mathcal{K}_\varepsilon$ if for any other test $\phi_T(\cdot) \in \mathcal{K}_\varepsilon$ and any $K > 0$ we have

$$\lim_{T\to\infty} \inf_{0\le\varrho\le K} \inf_{u(\cdot)\in\mathcal{U}_\varrho} \left[\beta_T(u, \phi_T^*) - \beta_T(u, \phi_T)\right] \ge 0.$$

Let us introduce the decision function

$$\hat{\phi}_T\left(X^T\right) = \chi_{\{\delta_T(X^T) > z_\varepsilon\}}, \qquad \delta_T\left(X^T\right) = \frac{X_T - S_* T}{\sqrt{S_* T}}.$$

Theorem 1.2.2 *Let $u(\cdot) \in \mathcal{L}^1(R_+) \cap \mathcal{L}^2(R_+)$, then the test $\hat{\phi}_T$ is locally asymptotically uniformly most powerful in the class $\mathcal{K}_\varepsilon$ and for any $u(\cdot) \in \mathcal{U}_\varrho$ its power function*

$$\beta_T\left(u, \hat{\phi}_T\right) \longrightarrow \hat{\beta}(u) = \mathbf{P}\left\{\zeta > z_\varepsilon - \varrho\sqrt{S_*}\right\},$$

where $\zeta \sim \mathcal{N}(0,1)$.

1.3 Self-Correcting Alternatives

We observe a trajectory $X^T = \{X_t, 0 \le t \le T\}$ of a point process of intensity function $S(\cdot)$ and there are two hypotheses: $\mathcal{H}_0: \quad S(t, X) = S_*$ and

$$\mathcal{H}_1: \quad S(t, X) = S_* \, \psi\left(\vartheta \left[S_* t - X_t\right]\right),$$

where $\psi(\cdot)$ is a known function.

Condition $\mathcal{A}$. *The function $\psi(x), x \in R$ is positive, continuously differentiable at the point $x = 0$, $\psi(0) = 1$ and $\dot{\psi}(0) > 0$.*

To have contigual alternative we put $\vartheta = u / S_* \dot{\psi}(0) T$. Therefore $\mathcal{H}_0: \ u = 0$ and $\mathcal{H}_1: \ u > 0$.

Denote

$$\Delta_T\left(X^T\right) \;=\; \frac{1}{S_* \, T} \int_0^T \left(S_* t - X_{t-}\right) \, \left[\mathrm{d}X_t - S_* \, \mathrm{d}t\right] = \frac{X_T - (X_T - S_* T)^2}{2 \, S_* T}.$$

and put $a_\varepsilon = \frac{1 - z^2_{\frac{1-\varepsilon}{2}}}{2}$ and $h(u) = \sqrt{\frac{2u}{1 - e^{-2u}}}$.

Theorem 1.3.1 *Let the Condition $\mathcal{A}$ be fulfilled, then the score function test $\phi^*_T\left(X^T\right) = \chi_{\{\Delta_T(X^T) > a_\varepsilon\}}$ belongs to the class $\mathcal{K}_\varepsilon$ and for any $u > 0$ its power function*

$$\beta_T(u, \phi^*_T) \to \beta^*(u) = \mathbf{P}\left\{|\zeta| \le h(u) \ z_{\frac{1-\varepsilon}{2}}\right\}.$$

The *likelihood ratio test* is

$$\bar{\phi}_T = \chi_{\left\{\delta_T(X^T) > \tilde{b}_\varepsilon\right\}}, \qquad \delta_T\left(X^T\right) = \sup_{\vartheta \in \Theta} L\left(\vartheta, X^T\right),$$

where $L(\vartheta, X^T)$ is the likelihood ratio function.

The *Wald's test*

$$\hat{\phi}_T\left(X^T\right) = \chi_{\left\{\gamma_T \hat{\vartheta}_T \ge c_\varepsilon\right\}}$$

where $\hat{\vartheta}_T$ is the maximum likelihood estimator of ϑ. The constants b_ε and c_ε are chosen from the condition $\bar{\phi}_T, \hat{\phi}_T \in \mathcal{K}_\varepsilon$.

We obtained the following presentation of the limit power functions of the score function test $\beta^*(u)$, likelihood ratio test $\bar{\beta}(u)$, Wald's test $\hat{\beta}(u)$ and Neyman-Pearson test β° :

$$\begin{aligned}
\beta^*(u) &= \mathbf{P}\left\{\int_0^u y_v \, \mathrm{d}w_v < \mathrm{J}_u - a_\varepsilon \, u\right\}, \\
\bar{\beta}(u) &= \mathbf{P}\left\{\int_0^u y_v \, \mathrm{d}w_v < \mathrm{J}_u - b_\varepsilon \sqrt{2\mathrm{J}_u}\right\}, \\
\hat{\beta}(u) &= \mathbf{P}\left\{\int_0^u y_v \, \mathrm{d}w_v < \mathrm{J}_u - \frac{c_\varepsilon}{u} \, \mathrm{J}_u\right\}, \\
\beta^\circ(u) &= \mathbf{P}\left\{\int_0^u y_v \, \mathrm{d}w_v < \frac{1}{2}\mathrm{J}_u + \frac{e_\varepsilon}{2} \, u^2\right\}.
\end{aligned}$$

where

$$\mathrm{d}y_v = -y_v\,\mathrm{d}v + \mathrm{d}w_v,\ \ y_0 = 0, \quad 0 \le v \le u, \qquad \mathrm{J}_u = \int_0^u y_v^2\,\mathrm{d}v.$$

Therefore for the large values of u

$$\tfrac{1}{2}\mathrm{J}_u + \tfrac{e_\varepsilon}{2}\,u^2 > \mathrm{J}_u - \tfrac{c_\varepsilon}{u}\,\mathrm{J}_u > \mathrm{J}_u - b_\varepsilon\,\sqrt{2\mathrm{J}_u} > \mathrm{J}_u - a_\varepsilon\,u,$$

and finally

$$\beta^*(u) < \bar{\beta}(u) < \hat{\beta}(u) < \beta^\circ(u).$$

Surprisingly the numerical simulation (with $N = 10^7$) shows practical coincidence of the last three powers. This effect is known in the time series statistics (AR process near singular point).

The proofs of the all results presented in this talk and the results of the simulation of the tests as well as the references of related works can be found on the site http://www.univ-lemans.fr/sciences/statist/index.php?page=publications of preprints of the University of Maine (see Preprints 05-3 and 05-4).

References

1. Cox, D.R. and Lewis, P.A.W. (1966). *Statistical Analysis of a Series of Events.* Methuen, London.

2. Dachian S., Kutoyants Yu.A. (2005). Hypotheses testing: Poisson versus self-exciting. Preprint 05-3, Université du Maine, (to appear in *Scandinavian J. of Statistics*)

3. Dachian S., Kutoyants Yu.A. (2005). Hypotheses testing: Poisson versus self-correcting. Preprint 05-4, Université du Maine,

4. Daley, D.J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. vol. I.* (2nd ed.), Springer, New York.

5. Hawkes, A.G. (1972). Spectra for some mutually exciting point processes with associated variable, In *Stochastic Point Processes.* (Ed., P.A.W. Lewis), Wiley, New York.

6. Isham, V. and Westcott, M. (1979). A self-correcting point processes, *Stochastic Processes and Applications*, **8**, 335–347.

7. Kutoyants, Yu. A. (1998) *Statistical Inference for Spatial Poisson Processes.* Springer, N.Y.

8. Pham, D.T. (1981). Estimation of the spectral parameters of a stationary point process, *Annals of Statistics*, **9**, 3, 615-627.

2

Asymptotic Confidence Bands For Kernel Density Estimators Based Upon Resampling

Paul Deheuvels

LSTA, Université Paris VI

Abstract: In this paper, we show that a single bootstrap suffices to construct sharp uniform asymptotic confidence bands for non-parametric kernel-type density estimators.

Keywords and phrases: Kernel density estimators, Non-parametric functional estimation, Bootstrap and resampling, Confidence bands.

2.1 Introduction and Results

Let $X_1, X_2, \ldots$ be a sequence of independent random replicæ of a random variable X with distribution function $F(x) = \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$. We are concerned with the estimation of the density $f(x) = \frac{d}{dx}F(x)$, assumed to exist, and to be continuous and positive on the interval $J = [c', d']$, where c' and d' are two constants such that $-\infty < c' < d' < \infty$. We will consider here the classical Akaike-Parzen-Rosenblatt (refer to ...) kernel estimator defined as follows. We first pick a kernel $K(\cdot)$, defined as a function of bounded variation on $\mathbb{R}$ such that

$$K(t) = 0 \quad \text{for} \quad |t| \geq \tfrac{1}{2} \quad \text{and} \quad \int_{\mathbb{R}} K(t)dt = 1.$$

We then select a bandswidth $h > 0$, and estimate $f(x)$ by the statistic

$$f_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big).$$

We are concerned with the limiting behavior of $f_{n,h}(x)$, uniformly over $x \in I = [c, d]$, where c and d are constants such that $c' < c < d < d'$. Setting $h_0 = \{c - c'\} \wedge \{d' - d\}$, we will assume, unless otherwise specified, that $h = h_n$

is a sequence depending upon n, and taking values within the interval $(0, h_0]$. The study of the uniform consistency of $f_{n,h}(x)$ to $f(x)$ makes use of the decomposition of $f_{n,h}(x) - f(x)$ into two components. The first one captures the *bias part*

$$\mathbb{E}f_{n,h}(x) - f(x) = \int_{-1}^{1} \{f(x - hu) - f(x)\}K(t),$$

which is easily shown to be independent of the sample size, $n \geq 1$, and to converge uniformly to 0 over $x \in I$, as long as $h \to 0$. The corresponding rate of convergence is a purely analytic problem, depending upon regularity assumptions on f, and will not be considered here. We will concentrate our interest on the *random part*

$$f_{n,h}(x) - \mathbb{E}f_{n,h}(x),$$

and investigate its uniform limiting behavior over $x \in I$. We seek *sharp asymptotic confidence bands*, defined as statistics θ_n, depending upon $n \geq 1$ and the sample $X_1, \ldots, X_n$, such that, as $n \to \infty$,

$$P\Big(\sup_{x\in I} |f_{n,h}(x) - \mathbb{E}f_{n,h}(x)| \leq \theta_n(1+\varepsilon)\Big) \to 1, \tag{2.1}$$

and

$$P\Big(\sup_{x\in I} |f_{n,h}(x) - \mathbb{E}f_{n,h}(x)| \leq \theta_n(1-\varepsilon)\Big) \to 0. \tag{2.2}$$

The limit law, stated in Fact 2.1.1 below, is due to Deheuvels and Einmahl (2000) (see also Stute (1982), Deheuvels (1992) and Deheuvels and Mason (1992)), and shows that a possible choice for θ_n in (2.1)–(2.8) is given by

$$\theta_n = \theta_{n,0} := \Big\{\frac{2\log(1/h_n)}{nh_n}\Big(\sup_{x\in I} f(x)\Big)\int_{\mathbb{R}} K^2(t)dt\Big\}^{1/2}. \tag{2.3}$$

Fact 2.1.1 *Let $\{h_n : n \geq 1\}$ be a sequence such that*

$(H.1)$ $\qquad h_n \to 0 \quad$ *and* $\quad nh_n/\log n \to \infty \quad$ *as* $\quad n \to \infty$.

Then, as $n \to \infty$,

$$\begin{aligned}&\sup_{x\in I} \pm\{f_{n,h_n}(x) - \mathbb{E}f_{n,h_n}(x)\}\\ &\qquad = (1 + o_{\mathbb{P}}(1))\Big\{\frac{2\log(1/h_n)}{nh_n}\Big\}^{1/2}\Big\{\sup_{x\in I} f(x)\int_{\mathbb{R}} K^2(t)dt\Big\}^{1/2}.\end{aligned} \tag{2.4}$$

Unfortunately, $\theta_{n,0}$, as given by (2.3), is not very useful, in practice, to construct limiting asymptotic bounds for $\sup_{x\in I} \pm\{f_{n,h_n}(x) - \mathbb{E}f_{n,h_n}(x)\}$. The fact that it depends upon the unknown density $f(\cdot)$ is a minor problem, since, as shown

in Deheuvels and Mason (2004), we may replace this quantity by $f_{n,h}(x)$ (or by another uniformly consistent estimator of $f(x)$), thus yielding

$$\theta_n = \theta_{n,1} := \Big\{ \frac{2\log(1/h_n)}{nh_n} \Big(\sup_{x\in I} f_{n,h}(x) \Big) \int_{\mathbb{R}} K^2(t)dt \Big\}^{1/2}. \tag{2.5}$$

The main difficulty, in the use of either $\theta_{n,0}$ or $\theta_{n,1}$, is due to the factor $\log(1/h_n)$, which is *scale-dependent*, up to the point where it becomes meaningless if the scale is chosen such that $h_n > 1$ for the value of n pertaining to the sample of interest.

The purpose of the present paper is to propose a simple and practical way to override the above-mentioned difficulty, based upon a resampling methodology. We start by the introduction of a sequence $\{Z_n : n \geq 1\}$ of independent and identically distributed random replicæ of a random variable Z. We assume that $\{X_n : n \geq 1\}$ and $\{Z_n : n \geq 1\}$ are independent. Moreover, we assume that the following conditions are satisfied.

$(A.1)$ $\quad \mathbb{E}(Z) = 1; \quad \mathbb{E}(Z^2) = 2$ (so that $\mathrm{Var}(Z) = 1$);
$(A.2)$ $\quad \mathbb{E}(e^{tZ}) < \infty$ for all $|t| < \epsilon$, for some $\epsilon > 0$.

We denote by $T_n = Z_1 + \ldots + Z_n$ the partial sum of order $n \geq 1$ of these random variables, and denote by $\mathcal{E}_n = \{T_n > 0\}$ the event that $T_n > 0$. We define, further, the random weights

$$\begin{aligned} W_{i,n} &= Z_i/T_n = \frac{Y_i}{\sum_{j=1}^n Y_j} \quad \text{for} \quad i = 1, \ldots, n \quad \text{when} \quad T_n > 0, \\ &= \frac{1}{n} \quad \text{when} \quad T_n \leq 0. \end{aligned} \tag{2.6}$$

We then define a *resampled* or *bootstrapped* version of $f_{n,h}(\cdot)$ by setting, for $h > 0$ and $x \in \mathbb{R}$,

$$f^*_{n,h}(x) = \frac{1}{h} \sum_{i=1}^n W_{i,n} K\Big(\frac{x - X_i}{h} \Big). \tag{2.7}$$

The following main result will turn out to provide a solution to our problem.

Theorem 2.1.1 *Under* $(H.1)$ *and* $(A.1$–$2)$*, we have, as* $n \to \infty$*,*

$$\begin{aligned} &\sup_{x\in I} \pm\{f^*_{n,h}(x) - f_{n,h_n}(x)\} \\ &\qquad = (1 + o_{\mathbb{P}}(1)) \Big\{ \frac{2\log(1/h_n)}{nh_n} \Big\}^{1/2} \Big\{ \sup_{x\in I} f(x) \int_{\mathbb{R}} K^2(t)dt \Big\}^{1/2}. \end{aligned} \tag{2.8}$$

This allows us to choose θ_n by setting

$$\theta_n = \theta_{n,2} := \sup_{x\in I} |f^*_{n,h}(x) - f_{n,h_n}(x)|. \tag{2.9}$$

2.2 Proofs

The proof of Theorem 2.1.1 relies on a version of Corollary 3.1 of Deheuvels and Mason (2004), stated in Fact 2.2.1 below. For the statement of this result, we will need the following notation. Denote by $G(z) = \mathbb{P}(Z \leq z)$ the distribution function of Z, and let $\psi(u) = \inf\{z : G(z) \geq u\}$, for $0 < u < 1$ denote the corresponding quantile function. Without loss of generality, it is possible to enlarge the original probability space $(\Omega, \mathcal{A}, \mathbb{P})$, in order to carry a sequence $\{Y_n : n \geq 1\}$ of independent and identically distributed random variables, with a uniform distribution on $(0, 1)$, and such that $Z_n = \psi(Y_n)$ for $n \geq 1$. Set now

$$r_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} \psi(Y_i) K\Big(\frac{x - X_i}{h}\Big). \tag{2.10}$$

We note that, under $(A.1)$, $\mathbb{E} r_{n,h}(x) = \mathbb{E} f_{n,h}(x)$. Therefore, it is readily checked from Corollary 3.1 of Deheuvels and Mason (2004), taken with $\lambda_1 = \lambda_2 = 1$, $c(x) = 1$ and $d(x) = 0$, that the following fact holds.

Fact 2.2.1 *Under* $(H.1)$ *and* $(A.1$–$2)$*, we have, as* $n \to \infty$*,*

$$\begin{aligned} &\sup_{x \in I} \pm\{r_{n,h_n}(x) - \mathbb{E} f_{n,h_n}(x)\} \\ &\qquad = (1 + o_{\mathbb{P}}(1)) \Big\{ \frac{2 \log(1/h_n)}{n h_n} \Big\}^{1/2} \Big\{ \sup_{x \in I} f(x) \int_{\mathbb{R}} K^2(t) dt \Big\}^{1/2}. \end{aligned} \tag{2.11}$$

Proof of Theorem 2.1.1. In view of (2.11), all we need for the proof of Theorem 2.1.1 is to show that, under $(H.1)$ and $(A.1$–$2)$, we have

$$\sup_{x \in I} |f_{n,h_n}(x) - r_{n,h_n}(x)| = o_{\mathbb{P}}\Big(\frac{\log(1/h_n)}{n h_n} \Big)^{1/2}. \tag{2.12}$$

Now, on the event $\mathcal{E}_n$ of (2.6), we have, by combining $(H.1)$ with the central limit theorem,

$$\begin{aligned} &\sup_{x \in I} |f_{n,h_n}(x) - r_{n,h_n}(x)| = \Big| \frac{n}{\sum_{i=1}^{n} \Psi(Y_i)} - 1 \Big| \sup_{x \in I} |r_{n,h_n}(x)| \\ &= O_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}\Big(\frac{\log(1/h_n)}{n h_n} \Big)^{1/2}. \end{aligned}$$

Since, obviously, $\mathbb{P}(\mathcal{E}_n) \to 1$, this, in turn, implies readily (2.12).□

Remark 2.2.1 *A similar, but slightly more involved argument, shows readily that the weights* $\{W(i, n) : 1 \leq i \leq n\}$ *may be chosen under the multinomial framework of the classical Efron bootstrap. Details about this will be given elsewhere.*

References

1. Deheuvels, P. (1992). Functional laws of the iterated logarithm for large increments of empirical and quantile processes, *Stochastic Processes and their Applications*, **43**, 133–163.

2. Deheuvels, P. and Mason, D. M. (1992). Functional laws of the iterated logarithm for the increments of empirical and quantile processes, *Annals of Probability*, **20**, 1248–1287.

3. Deheuvels, P. and Mason, D. M. (2004). General asymptotic confidence bands based upon kernel-type function estimators, *Statistical Inference for Stochastic Processes*, **7**, 225–277.

4. Stute, W. (1982). The oscillation behavior of empirical processes, *Annals of Probability*, **10**, 414–422.

3

Frailty Models for Censored and Truncated Data

Catherine Huber-Carol, Valentin Solev and Filia Vonta

Université Paris V, René Descartes, 45 rue des Saints-Pères, 75 006 Paris, and INSERM U 780.
Steklov Institute of Mathematics at St. Petersburg, nab. Fontanka, 27 St.Petersburg 191023 Russia
Department of Mathematics and Statistics, University of Cyprus P.O. Box 20537, CY-1678, Nicosia, Cyprus

Abstract: Semi-parametric estimation of survival data with censoring and truncation under a frailty model with covariates is proposed, based on a maximum likelihood approach. Consistency of the NPML estimate is proved when no covariates are present.

Keywords and phrases: Frailty, censoring, truncation, identifiability, semi-parametric estimation, consistency

3.1 NPML estimation for frailty model

3.1.1 The model

X_i is a positive random variable, the survival of subject i, $i = 1, \cdots, n$, whose p-dimensional covariate z_i is observed and frailty η_i is not observed, but has known distribution function F_η on $\mathbb{R}^+$. The $X_i's$ are independent and their survival function $S(t) = P(X \geq t)$ and hazard function $h(t) = -dS(t)/S(t)$ obey the following frailty model, where $\beta \in R^p$ is an unknown regression parameter, $\lambda(t)$ the unknown baseline hazard and $\Lambda(t) = \int_0^t \lambda(u)du$ the corresponding cumulative baseline hazard:

$$\begin{aligned} h(t|z,\eta) &= \eta e^{\beta^T z}\lambda(t) \qquad (3.1)\\ S(t|z,\eta) &= e^{-\eta e^{\beta^T z}\Lambda(t)} \end{aligned}$$

$$S(t|z) = \int_0^\infty e^{-xe^{\beta^T z}\Lambda(t)} dF_\eta(x) = e^{-G(e^{\beta^T z}\Lambda(t))} \qquad (3.2)$$

where G is equal to $-\log$ of the Laplace transform of η:

$$G(y) = -\ln(\int_0^\infty e^{-uy} dF_\eta(u)) \qquad (3.3)$$

3.1.2 Censoring and truncation

X may be both censored and truncated so that the $X_i's$ are generally not observed. Instead, one observes two intervals (A_i, B_i), which are respectively the censoring interval, $A_i = [L_i \ ; \ R_i]$, and the truncating interval $]\mathcal{L}_i \ ; \ \mathcal{R}_i[$, such that $B_i \supset A_i$. This means that X_i is not observed but is known to lie inside A_i, and A_i itself is observed only conditionally on the fact that it is inside the truncating interval B_i. Otherwise, the corresponding subject is said to be "truncated" i.e. it does not appear in the sample. Finally, for the n subjects which are not truncated, the observations are $(A_i, B_i, z_i), i \in \{1, 2, \cdots, n\}$.

3.1.3 The likelihood

The likelihood is proportional to

$$l(S) = \prod_{i=1}^{n} l_i(S_i) = \prod_{i=1}^{n} \frac{P_{S_i}(A_i)}{P_{S_i}(B_i)} = \prod_{i=1}^{n} \frac{\{S_i(L_i^-) - S_i(R_i^+)\}}{\{S_i(\mathcal{L}_i^+) - S_i(\mathcal{R}_i^-)\}} \tag{3.4}$$

Following Turnbull (1976), let us define the "beginning" set $\tilde{L}$ and "finishing" set $\tilde{R}$, in order to take advantage of the fact that the likelihood is maximum when the values of $S_i(x)$ are the greatest possible for $x \in \widetilde{L}$ and the smallest possible for $x \in \widetilde{R}$:

$$\begin{aligned} \widetilde{L} &= \{L_i, 1 \le i \le n\} \cup \{\mathcal{R}_i, 1 \le i \le n\} \cup \{0\} \\ \widetilde{R} &= \{R_i, 1 \le i \le n\} \cup \{\mathcal{L}_i, 1 \le i \le n\} \cup \{\infty\}. \end{aligned}$$

Let

$$Q = \{[q'_j, p'_j] : q'_j \in \widetilde{L} \ , \ p'_j \in \widetilde{R} \ , \ [q'_j, p'_j] \cap \widetilde{L} = \varnothing \ , \ [q'_j, p'_j] \cap \tilde{R} = \varnothing\}$$

$$0 = q'_1 \le p'_1 < q'_2 \le p'_2 < \ldots < q'_v \le p'_v = \infty$$

Then,

$$Q = \cup_{j=1}^{v} [q'_j, p'_j] = C \cup W \cup D$$

where

$$\begin{array}{rcll} C & = & \cup[q'_j, p'_j] & \text{covered by at least one censoring set,} \\ W & = & \cup[q'_j, p'_j] & \text{covered by at least one truncating set,} \\ & & & \text{but not covered by any censoring set,} \\ D & = & \cup[q'_j, p'_j] & \text{not covered by any truncating set.} \end{array}$$

The special case with G, defined in (1.3), being the identity function, and $\beta = 0$ was studied in detail in Turnbull (1976), Frydman (1994) and Finkelstein et. al (1993).

The above likelihood, as a function of the unknown β and Λ, is equal to

$$l(\Lambda,\beta|(A_i,B_i,z_i)_{i\in\{1,..,n\}}) = \prod_{i=1}^{n} \frac{\left\{e^{-G(e^{\beta^T z_i}\Lambda(L_i^-))} - e^{-G(e^{\beta^T z_i}\Lambda(R_i^+))}\right\}}{\left\{e^{-G(e^{\beta^T z_i}\Lambda(\mathcal{L}_i^+))} - e^{-G(e^{\beta^T z_i}\Lambda(\mathcal{R}_i^-))}\right\}} \quad (3.5)$$

3.1.4 NPML estimation

As in the case of Turnbull (1976), where G is the identity and $\beta = 0$ in (3.2), the NPML estimator of Λ for the frailty model (3.1) is not increasing outside the set $C \cup D$ (Huber and Vonta, 2004). Also, the set C is written as $C = \cup_{j=1}^{m}[q_j, p_j]$. Moreover, conditionally on the values of $\Lambda(q_j{}^-)$ and $\Lambda(p_j{}^+)$, $1 \le j \le m$, the likelihood does not depend on how the mass $\Lambda(p_j{}^+) - \Lambda(q_j{}^-)$ is distributed in the interval $[q_j, p_j]$. The set D may be expressed as $D = \cup_{j=0}^{m} D_j$, where $D_j = D \cap (p_j, q_{j+1})$, $p_0 = 0$, $q_{m+1} = \infty$. Let $\delta_j = P_\Lambda(D_j)$ and the indicators $\mu_{ij} = I\{[\ [q_j, p_j] \subset A_i]\}$ and $\nu_{ij} = I\{[\ [q_j, p_j] \subset B_i]\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Then the log-likelihood is equal to

$$\begin{aligned}
\log l(\Lambda,\beta|z_1,\ldots,z_n) \;=\; & \textstyle\sum_{i=1}^{n}\Big\{ \log\Big(\sum_{j=1}^{m} \mu_{ij} \quad \big(e^{-G(e^{\beta^T z_i}(\Lambda(p_{j-1}^+)+\delta_{j-1}))} \\
& \qquad\qquad -e^{-G(e^{\beta^T z_i}\Lambda(p_j{}^+))}\big)\Big) \\
& - \log\Big(\textstyle\sum_{j=1}^{m} \nu_{ij} \qquad\quad \big(e^{-G(e^{\beta^T z_i}(\Lambda(p_{j-1}^+)+\delta_{j-1}))} \\
& \qquad\qquad -e^{-G(e^{\beta^T z_i}\Lambda(p_j{}^+))}\big)\Big)\Big\}.
\end{aligned}$$

In most practical cases $D = D_0$ (left truncation) and/or $D = D_m$ (right truncation). In order to ensure the positivity of the estimators of $\delta's$ and $\Lambda's$ and the monotonicity of the sequence of $\gamma's$ below, we change parameters:

$$\begin{array}{lllllll}
\gamma_0 & = & \log(\delta_0) & \gamma_j & = & \log(\Lambda(p_j)) & j = 1,\ldots,m \\
\tau_1 & = & \gamma_1 & \tau_j & = & \log(\gamma_j - \gamma_{j-1}) & j = 2,\ldots,m.
\end{array}$$

We want now to maximize:

$$\begin{aligned}
\log l(\Lambda,\beta|z) = \sum_{i=1}^{n}\Big\{ & \log\Big(\sum_{j=1}^{m} \mu_{ij}\big(e^{-G(e^{\beta^T z_i+\tau_1+\sum_{k=2}^{j-1} e^{\tau_k}})} - e^{-G(e^{\beta^T z_i+\tau_1+\sum_{k=2}^{j} e^{\tau_k}})}\big)\Big) - \\
& \log\Big(\sum_{j=1}^{m} \nu_{ij}\big(e^{-G(e^{\beta^T z_i+\tau_1+\sum_{k=2}^{j} e^{\tau_k}})} - e^{-G(e^{\beta^T z_i+\tau_1+\sum_{k=2}^{j} e^{\tau_k}})}\big)\Big)\Big\}. \quad (3.6)
\end{aligned}$$

There is though a problem of identifiability as we have not any observation on D, which is not covered by any truncating set. It can be proved that when no covariate is present, what is identifiable is only the ratio $P([q_j; p_j])/\sum_{j=1}^{m} P([q_j; p_j])$,

for all $j \in \{1, \ldots, m\}$. When a covariate of dimension $p \geq 2$ is present, then it can be proved that there is identifiability, by showing that those parameters are functions of quantities that are themselves identifiable using the simple case of $D = D_0 \bigcup D_m$ and of two binary covariates leading to four different S_i and sets C and D that can be indexed by these four categories.

For simulations and a real data example (Kalbfleish AIDS data from transfusion) we used (Huber and Vonta, 2004), the two most popular frailty distributions: the Inverse Gaussian with mean 1 and variance $1/2b$, and the Gamma (Clayton-Cuzick frailty model) with mean 1 and variance c. The function G (3.3) takes respectively the form:

$$\begin{aligned} G(x,b) &= \sqrt{4b(b+x)} - 2b,\ b > 0 \\ G(x,c) &= \tfrac{1}{c}\ln(1+cx),\ c > 0. \end{aligned}$$

3.2 Consistency

We now assume that there is no covariate involved. We want to prove the consistency of the NPML estimator of Λ. The likelihood we maximized is *conditional on the censoring and truncating sets*. It does not take into account the laws of the censoring and truncation while the consistency is highly dependent on the features of those two laws that may or may not allow a precise non parametric estimation of Λ.
Let $(a, b) \subseteq I\!R^+$ be the interval where all observations take place. Let us define the censoring and truncating scheme on a practical example. In order to fix ideas, let us say that a patient is visited by a doctor at several dates $\tau = \{(Y_j, Y_{j+1}],\ j = 0, 1, \ldots, J(i)\}$, within a predetermined period of time $B =]Z_1; Z_2[\subseteq (ab)$, and the doctor observes that the expected event, for example AIDS onset, took place after Y_k and before or else at Y_{k+1}.

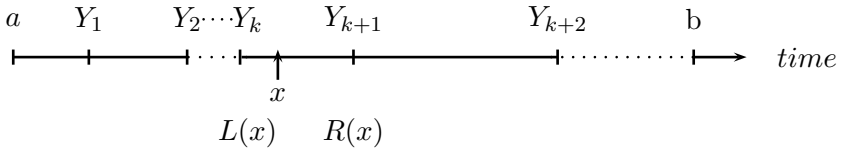


figure 1 : censoring interval]L(x) R(x)].

Then we know that x, the time to onset of the event is censored by $]L\,;\,R] = (Y_k Y_{k+1}]$ and truncated by $(\mathcal{L}\,;\,\mathcal{R}) = (Z_1; Z_2)$ or more precisely by $[Y_1\ ;\ Y_{J(i)}]$, the "*observed* truncating set". To each subject i, there corresponds a "random covering" τ_i, a truncating pair $Z = (Z_{i1}, Z_{i2})$ and a survival X_i. The three random elements are assumed to be independent, and i.i.d. when i varies from 1 to n.

Dropping now for simplicity the subscript i, let τ be a random covering and define, for every u in $\mathbb{R}^+$, $L(u)$ as the greatest element of the covering τ that is smaller than u and $R(u)$ the smallest element of τ that is greater than or equal to u. More precisely $(L(u), R(u)) = (Y_k, Y_{k+1})$, where $k = k(u) = \inf\{j : u \leq Y_{j+1}\}$. Then, for the subject associated to this covering, with survival x and truncating set $(z_1; z_2)$, applying this to $u = x, u = z_1$ and $u = z_2$, we have two cases: either $R(z_1) \leq L(x) \leq R(x) \leq L(z_2)$, then we observe the pair of censoring and truncating sets $(A, B) := (]L(x)\ ;\ R(x)], [R(z_1)\ ;\ L(z_2)])$, or $R(z_1) > L(z_2)$, then the subject is truncated and gives no observation. The two cases are illustrated in figure 2 below: above the time line, $L(z_2) < R(z_1)$, so that x is truncated and there is no observation, while below the line, the censoring $]L(x); R(x)]$ and truncating $[R(z_1); L(z_2)]$ sets are both observed as $(R(Z_1) < L(X) < R(X) < L(Z_2))$.

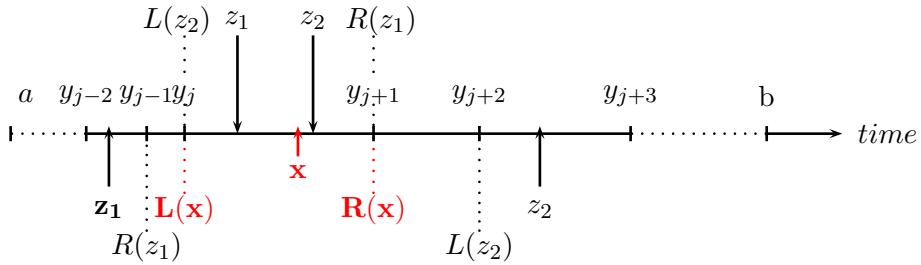


figure 2 : two examples of truncating sets on the same x and τ.

Let P_x be the law of $v(x) = (L(x), R(x))$. If for all x, $P_x \ll \lambda^2$, with density $r_x(u, v) = \frac{dP_x}{d\lambda^2}$, then there exists a non-negative function $r(u, v)$, called the *baseline density of the random covering*, such that for all x

$$r_x(u, v) = r(u, v)\ \mathbb{1}_{(u,\, v]}(x)\ \text{(a.s.)} \tag{3.7}$$

A necessary and sufficient condition for $P_x \ll \lambda^2$ to hold for all x is that
(i) for all j , $dP_{(Y_j, Y_{j+1})}/d\lambda^2 = r^j(u, v)$,
(ii) The series $\sum_j r^j(u, v)$ converges a.s. to a function $r(u, v)$,
(iii)For all x, $r_x(u, v) = r(u, v)\ \mathbb{1}_{(u,\, v]}(x)$.
For a proof and details, see Huber, Solev and Vonta (2005). Adding the truncation, we now deal with the law P_{x,z_1,z_2} of vector $W = (R(Z_1), L(X), R(X), L(Z_2))$. This law has three components. One is AC with respect to λ^4, when $R(Z_1) < L(X) < R(X) < L(Z_2))$, one is AC with respect to λ^3, when $R(Z_1) = L(X) < R(X) < L(Z_2)$ or $R(Z_1) < L(X) < R(X) = L(Z_2)$, and one is AC with respect to λ^2, when $R(Z_1) = L(X) < R(X) = L(Z_2)$. Denoting ν the baseline measure having density 1 with respect to the three Lebesgue measures, let $q_{x,z} = dP_{x,z_1,z_2}/d\nu$. Let us now assume that, for all $n, m > 0$, $P_{Y_n,\ldots,Y_{n+m}} \ll \lambda^{m+1}$, and let us denote for $i + 1 < j$:

$$
\begin{array}{lll}
r_{i,j}(y_1,y_2,y_3,y_4) & \text{the density of} & (Y_i,Y_{i+1},Y_j,Y_{j+1}),\\
r_j(y_1,y_2,y_3) & \text{""} & (Y_{j-1},Y_j,Y_{j+1}),\\
r^j(y_1,y_2) & \text{""} & (Y_j,Y_{j+1}).
\end{array}
$$

We assume convergence of the series

$$
\begin{array}{lcll}
\partial_4(s_1,s_2,s_3,s_4) & = & \sum\limits_{i,\,j:i+1<j} r_{i\,j}(s_1,s_2,s_3,s_4)<\infty & (\lambda^4\text{-a.s.}),\\
\partial_3(s_1,s_2,s_3) & = & \sum\limits_{j} r_j(s_1,s_2,s_3)<\infty & (\lambda^3\text{-a.s.}),\\
\partial_2(s_1,s_2) & = & \sum\limits_{j} r^j(s_1,s_2)<\infty & (\lambda^2\text{-a.s.}).
\end{array}
$$

Then $q_{x,z}(s_1,s_2,s_3,s_4) = \mathbb{1}]s_2,s_3](x)\{\partial_2(s_2,s_3)\ \mathbb{1}_{[s_2,s_3]}(z)+\partial_3(s_1,s_2,s_3)\ \mathbb{1}_{[s_3,s_4]}(z)+\partial_4(s_1,s_2,s_3,s_4)\ \mathbb{1}_{[s_1,s_4]}(z)$. If now we assume to simplify without restriction of the generality, that there is only right truncation, by setting $z_1=a$ and denoting $L(z_2)=z$, then ∂_2 disappears in the above formula, and as $R(z_2)$ plays no role, integration with respect to s_4 can take place so that one finally gets the following formula for the law p_w of the observation W:

$$
\begin{array}{lcl}
p(u,v,z) & = & r(u,v,z)\times \dfrac{\int_u^v f(x)\,dx}{\int_{x\le z} f(x)\,dx}\\[2ex]
& = & r(u,v,z)\times\varphi(f|u,v,z)
\end{array}
$$

$$
r\in\mathcal{G}\ ,\qquad f\in\mathcal{F}\quad ,\ p\in\mathscr{P}=\{r\,\varphi(f|\cdot),\ (r,f)\in\mathcal{G}\times\mathcal{F}\}
$$

where f is the density of the survival X and r is the density of the censoring-truncation based on the random covering. The NPML estimate $\widehat{p_n}$, of $p\in\mathcal{P}$ is defined as $\widehat{p_n}=Arg\max\limits_{q\in\mathcal{P}}\int \ln q\ dP_n$. ML estimates of $r\in\mathcal{G}$ and $f\in\mathcal{F}$ verify $\widehat{r_n}=Arg\max\limits_{q\in\mathcal{G}}\int \ln q\ dP_n$, $\widehat{f_n}=Arg\max\limits_{q\in\mathcal{F}}\int \ln\varphi(q)\ dP_n$, and $\widehat{p_n}=\widehat{r_n}\times\varphi(\widehat{f_n}|\cdot)$. The two results on which the proof of consistency of NPLM estimate of f relies are a lemma of Sara van de Geer (1993) and a sufficient condition for having a uniform law of large number in a class of functions (van der Vaart and Wellner (2000)). Sara Van de Geer lemma tells us that $h^2(\widehat{p_n},p_0)\le\int_{p>0}(\sqrt{\frac{\widehat{p_n}}{p_0}}-1)\,d(P_n-P_0)$. Denoting $\mathscr{P}^*:=\{(\sqrt{\frac{p}{p_0}}-1)\times\ \mathbb{1}\{p_0>0\}\ :p\in\mathscr{P}\}$, one sees that $h(\widehat{p_n},p_0)\to 0\ a.s.$ as soon as $\sup_{p^*\in\mathscr{P}^*}|\int\ p^*\,d(P_n-P_0)|\to 0\ \ a.s.$ So a necessary and sufficient condition to have consistency of $\widehat{p_n}$ is that we have a uniform law of large numbers on $\mathscr{P}^*$. This is satisfied in particular if $\mathscr{P}^*$

and its envelop are in $L^1(P_0)$ and their bracketing entropy $H_{[\,]}(\varepsilon, \mathscr{P}^*, L^1(P_0))$ is finite for all ε. The following assumptions ensure the consistency of NPML estimate of f:

C$_1$ The set $\mathscr{P}$ is totally bounded in Hellinger distance.

C$_2$ For a constant $C = C_{\mathscr{P}}$ and $\varepsilon > 0$ there exist finite coverings of $\mathcal{F}$, $\mathcal{G}$ and $\mathscr{P}$: $\mathcal{F} \subset \bigcup_{i=1}^{m} V(f_i^L, f_i^R)$, $\mathcal{G} \subset \bigcup_{j=1}^{k} W(r_j^L, r_j^R)$; $\mathscr{P} \subset \bigcup_{i,j} U(p_{i,j}^L, p_{i,j}^R)$ such that $\{p : p = p^{r,f}, \text{ for } r \in W(r_j^L, r_j^R) \text{ and } f \in V(f_i^L, f_i^R)\} \subset U_{i,j} = U(p_{i,j}^L, p_{i,j}^R)$;

C$_3$ $h(p_{i,j}^L, p_{i,j}^R) \le \varepsilon$, $h(f_i^L, f_i^R) \le \varepsilon$; $\int p_{i,j}^R \, d\mu < C_{\mathscr{P}}$, $\int f_i^R \, dx < C_{\mathscr{P}}$;

C$_4$ For all $\varepsilon > 0$, $z_0 < b + \varepsilon$: $\inf_{r \in W_j} \int_{v-u \le \varepsilon,\, z \ge z_0} r(u, v, z) \, d\mu > 0$.

References

1. Finkelstein, D. M., Moore, D. F. and Schoenfeld, D. A. (1993). A proportional hazards model for truncated AIDS data. *Biometrics* **49**, 731–740.

2. Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society, Series B* **56**, 71–74.

3. Huber-Carol, C. and Vonta, F. (2004). Semiparametric Transformation Models for Arbitrarily Censored and truncated Data, In *Parametric and Semiparametric Models with Applications to reliability, Survival Analysis, and Quality of Life* (Ed. M.S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios), pp. 167–178, Birkhauser, Boston.

4. Huber-Carol, C. Solev, V. and Vonta, F. (2005). Estimation of Density for Arbitrarily Censored and Truncated Data, In *Probability, Statistics and Modelling in Public Health* (Ed. Mikhail Nikulin, D. Comenges and C. Huber), pp. 246–265, Springer Verlag.

5. Turnbull, B.W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data, *Journal of the Royal Statistical Society*, **38**, 290–295.

6. Van de Geer Sara (1993). Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators, *The Annals of Statistics*, **21**, 14–44.

7. Van der Vaart, Aad W. and Wellner Jon A. (2000). *Weak Convergence and Empirical Processes. With Applications to Statistics*, Springer Verlag, 2nd edition.

4

Clinical Trials and the Genomic Evolution: Some Statistical Perspectives

Pranab Kumar Sen

University of North Carolina, USA

4.1 Extended Abstract

A little over a period of three decades, clinical trials have mushroomed in a variety of human health studies, with a variety of objectives, having a variety of interdisciplinary perspectives, and diverse implementational motives. Clinical trials are designed by human beings, mostly, for human beings, incorporating mostly human subjects, and, supposedly, for human benefit. Yet in this human venture there are some inhuman features which warrant critical appraisal. Using human subjects in scientific (and mostly exploratory) studies may generally trigger *medical ethics, cost-benefit perspectives* and a variety of other concerns. In order to control some of these disturbing concerns, often, subhuman primates are advocated as precursors or surrogates of human being, albeit there remains a basic query:
How to extrapolate stochastics from mice to man? Can the basic principles of animal studies or dosimetry be validated in clinical trials designated for human being?
There is a basic qualm on the main objective of a clinical trial: *symptomatic effects* versus true disease-disorder detection and cure. Drug developers, pharmaceutical groups and regulatory agencies focus on treatments to relieve symptoms which may not totally or adequately match treatment objectives. Bioethics and public advocates have voiced concern on clinical trials in third-world countries, the affordibality of usual *high-cost drugs* being a major issue in this cost-benefit context. WHO and public health authorities all over the world are trying to identify effective and affordable regimens for many developing countries. These medical ethics, economic resources (affordability) and operational restraints often mar the routine use of standard statistical tools for drawing valid conclusions from clinical trials.

There are some basic differences between animal studies and clinical trials. The former can be conducted in a fairly controlled laboratory setups but human beings can not be put under such controlled environments, and as such, the enormous disparity in physical characteristics and many other epidemiologic endpoints, call for highly nonstandard statistical modeling and analysis. That is why *placebo-controlled trials* (PCT) are used extensively in development of new pharmaceuticals. In the early phase of adoption of clinical trials, such PCTs were mostly advocated. However, there are allegations that PCT are invariably unethical when known effective therapy is available for the condition being treated or studied, regardless of the condition or the consequences of deferring treatments. The 1997 Helsinki Declaration by the World Medical Association (WMA) has clearly laid down the ethical principles for clinical trials: *In any medical study, every patient - including those of a control group, if any, should be assured of the best proven diagnostic and therapeutic methods.* Most often, in a PCT, this ethics is violated by the very composition of the placebo group. Based on this declaration, patients asked to participate in a PCT must be informed of the existence of any effective therapy, must be able to explore the consequences of deferring such therapy with the investigator, and must provide fully informed consent. *Active controlled equivalence trials* (ACET) have therefore been advocated for comparing an existing treatment with a targeted one. They may show whether a new therapy is superior (or inferior) to an existing one, but may not possess other characteristics of PCTs (Temple and Ellenberg 2000, Sen 2001).

No matter it is a PCT or an ACET, there are numerous underlying constraints calling for novel *constrained statistical inference* (CSI) tools for statistical analysis. There is another feature common to both PCT and ACETs. It may be desirable in such a follow-up study to have *interim analysis* to monitor the accumulating clinical evidence in the light of statistical perspectives. While this feature has led to the evolution of *time-sequential* statistical methodology, there remains much to update this novel branch of CSI in the light of the underlying constraints and complications. It is usually desirable to look into the accumulating data sets at regular time intervals, and statistically deciding whether or not an *early termination* of the trial can be made in favor of the new therapy (if that is to be advocated in the drug market) so that patients can be switched to a better health perspective. Thus, usually, a *repeated significance testing* (RST) scheme, often in a restrained setup, underlies statistical modeling and analysis of clinical trials. In conventional *group sequential tests* (GST) usually one assumes independent and homogeneous increments for the associated stochastic processes. This is generally not the case in interim analysis related RST. *Progressively censoring schemes* (PCS) were introduced by Chatterjee and Sen (1973) to formulate the general methodology of *time-sequential* procedures; suitable martingale characterisations underlie most of these developments (Sen

1981, 1999a, 2001). With a need to update this approach in a more general framework to suit the ACET, let us consider the following statistical scenario.

Consider a typical constrained statistical interim analysis scheme relating to a comparative clinical trial relating to an existing therapy and a new one. The interim analysis related to monitoring of the accumulating evidence at time points $t_1 < \cdots < t_K$ for some specified K, spanning over a projected period of study $T = (0, t_K)$. If at an early time point t_k, there appears to be a significant difference (in favor of the new drug), then the trial is to be terminated at that point. The null hypothesis (H_0) relates to no difference over the entire period T while the alternative (H_1) to the new being better than the existing. We frame the null hypothesis H_{0r} that up to the time point t_r there is no difference between the two therapies, and let H_{1r} be the alternative that for the first time, at time point t_r, there is a difference in favor of the new drug, for $r = 1, \ldots, K$. Then, restricted to the time domain T, we may note that there there is a nested nature of these hypotheses. The null hypothesis H_0 is accepted only when all the H_{0r} are accepted, while the alternative hypothesis H_1 is accepted when at least one of the K exclusive hypotheses $H_{1r}, 1 \leq r \leq K$ is accepted. Hence we write

$$H_0 = \cap_{r=1}^{K} H_{0r}, \quad H_1 = \cup_{r=1}^{K} H_{1r}. \tag{4.1}$$

Further, based on the accumulating data set upto the time point t_r, we construct a suitable test statistic $\mathcal{L}_r$ for testing H_{0r} vs $H_{1r},\ r = 1, \ldots, K$. This is essentially a RST problem in a constrained environment, and the natrure of the null and alternative hypotheses immediately calls for the *union intersection principle* (UIP). There is, however, some notable differences between the clinical trial and usual multiple hypothesis testing problems. The UIP having a finite intersection / union composition is more cumbersome to incorporate. Because of clinical and ethical undercurrents, first we appraise the potential constraints.

Restraint 1 : The component hypotheses are nested. For each $r(= 1, \ldots, K,\ H_{1r}$ is a one-sided alternative.

Restraint 2 : For different $r(= 1, \ldots, K)$, the different test statistics $\mathcal{L}_r$ are not independent, and the pattern of their dependence may not follow a Markov chain.

Restraint 3 : Early termination of the trial is associated with the acceptance of H_{1r}, for some $r < K$. It might be also due to significant adverse side-effects of the treatment, irrespective of the accumulating statistical evidence.

Restraint 4 : Explanatory variables provide useful statistical information, and hence, need to be included as far as possible, albeit increasing model complexity and CSI protocols.

Restraint 5 : Conventional (log-)linear regression models may not be appropriate. Some of the explanatory variables (viz., smoking, physical exercise, diabetic, etc.) may be binary, or at best, categorical. Even if they were quantitave, often for data recording, they are are reported as categorical.

Restraint 6 : Informative censoring: Censoring due to noncompliance (e.g., drop-out or failure due to other causes) may not be independent of the placebo-treatment setup.

Restraint 7 : Surrogate end point: Often, the primary end point may be costly from data collection perspectives, and some closely related or associated (by symptoms, for example) variables, termed surrogate end points are used as substitute. The statistical model for the surrogate end point could be quite different from the primary one. Further, multiple end points may also crop up in such studies. Standard parametric multivariate CSI tools may not be usable properly.

Restraint 8 : Assessment of statistical quality of accumulating data with due respect to the underlying clinical and statistical restraints could be a major task.

Restraint 9 : Parametric models may not suit the purpose. Nonparametrics and semiparametrics may perform better. However, the underlying restraints in semiparametrics may generally need critical appraisal. Nonparametrics may fare better but may require larger sample sizes to be of good quality and efficacy.

Restraint 10: Data mining : The advent of genomics is increasingly advocating for large number of end points and explanatory variables, and knowledge discovery and data mining (KDDM) tools are being advocated more and more. This does not, however, diminish the primary concern: To what extent statistical inference is not compromised or invalidated by data mining?

Suppose now that taking into account most of these restraints, albeit in approximate forms, it is possible to observe the partial data set $\mathcal{D}_t$ upto the time point t, so that $\mathcal{D}_t$ is nondecreasing (accumulating) in $t \in T$. Let $\mathcal{F}_t$ be the history process upto the time point t, so that $\mathcal{F}_t$ is nondecreasing in $t \in T$. Further, suppose that if all the (n)observations were available (i.e., the data set includes all responses and all explanatory variables), then for testing H_0 against a restricted alternative H_1, we would have a desirable test statistic which we denote by $\mathcal{L}_n$. In a parametric setup, $\mathcal{L}_n$ could be a version of the likelihood ratio statistic or some of its variants like the partial-, penalized likelihood score etc. In semiparametrics, pseudo-, quasi-, or profile likelihood statistics might

be usable. In nonparametrics, rank statistics have more appeal. We may set without any loss of generality $E\mathcal{L}_n|H_0) = 0$, and define

$$\mathcal{L}_n(t) = E_{H_0}\{\mathcal{L}_n \mid \mathcal{F}_t\}, t \geq 0. \tag{4.2}$$

Then, under fairly general regularity assumptions, we may note that under H_0,

$$\{\mathcal{L}_n(t), \mathcal{F}_t; t \geq 0\} \text{ is a zero mean martingale (array)}, \tag{4.3}$$

although this martingale characterization may not generally hold when the null hypothesis is not true. Also, even under the null hypothesis, $\mathcal{L}_n(t)$ may not have independent and stationary increments. Our task is to set a time sequential or RST procedure based on the discretized time-parameter process $\{\mathcal{L}_n(t_j),\ j \leq K\}$. Thus, we are confronted with suitable CSI procedures amenable to RST or interim analysis. Intuitively, we could conceive of an array of cut-off points: $\{C_{nr}, r = 1l, \ldots, K\}$, such that if $\mathcal{L}_n(t_1) \geq C_{n1}$, we stop the trial along with the rejection of H_0; if not, we go to the next time period t_2 and then if $\mathcal{L}_n(t_2) \geq C_{n2}$, we stop at that time along with the rejection of the null hypothesis. Otherwise we proceed to the next time period. In this way, the process continues, and if for the first time, for some $k \leq K,\ \mathcal{L}_n(t_k) \geq C_{nk}$, we reject the null hypothesis at that point and stop the trial. Thus, we proceed to accept the null hypothesis only when $\mathcal{L}_n(t_j) < c_{nj}, \forall j \leq K$, continuing the trial to its target time t_K.

The basic problem is to control the overall Type I error rate without sacrificing much power in such interim analyses scheme. This, in turn, requires a skillful choice of the cut-off points $C_{nr}, r \leq K$, which generally depend not only on the $t_k, k \leq K$ but also on the accumulated statistical information at these points, and the latter is generally unknown or, at least, not properly estimable at the start of the trial. In this respect, we shall appraise the role of UIP along with other competitors. Group sequential tests (GST), formulated mostly in the late 1970's, make explicit use of normal distribution and equal increment assumptions which may not be generally true in such a time sequential setup. Even so, they needed extensive computation of the cut-off points. For some of these details, we refer to Sen (1999). Led by the basic weak convergence results for progressively censored linear rank statistics (Chatterjee and Sen 1973) some of these computational complexities have been eliminated considerably.

Typically, there exists a (random) time-parameter transformation by which the process $\{\mathcal{L}_n(t), t \in T\}$ can be written as $\mathbf{W}_{n,T} = \{W_{n,T}(u), u \in [0,1]\}$ such that under the null hypothesis, $\mathbf{W}_{n,T}$ converges weakly to a Brownian motion on $[0,1]$. By the same transformation, the calendar time points $t_r, r = 1, \ldots K$ are converted into (random) information time points $u_1 < \cdots < u_K$. Thus, we reduce the problem to a multivariate one-sided alternative hypothesis testing CSI problem for which the UIT sketched in earlier sections works out well.

Basically, we have to construct the $W_{n,T}(u_r), r \geq 1$, and find a suitable cut-off point τ_{α^*} and a significance level α^* such that for a chosen α,

$$P\{W_{n,T}(u_r)/\sqrt{u_r} \ < \tau_{\alpha^*}, \forall r \ | H_0\} \leq \alpha. \tag{4.4}$$

Since a Brownian motion process $W(t), t \in [0,1]$ has irrerular behavior with respect to the square root boundary as $t \to 0$, technically, we need that u_1 is away from 0. If the u_r are scattered over $(0,1]$ and K is large, a more convenient way of computing the cut-off points would be to appeal to the boundary crossing probability of standard Brownian motion over one-sided square root boundaries; DeLong (1981) has provided detailed tables for these. This approximation is quite good when K is larger than 10, as is often the case of clinical trials with long-range follow-up time. Here also, the tabulated critical values correspond to some small truncation at 0 (i.e., over the range $[\epsilon, 1]$, for some positive ϵ (small). This weak invariance principle also avoids the need to specify the exact information times needed for the GST. There is an allied RST procedure considered by Chatterjee and Sen (1973) [and DeMet and Lan (1983)] where the weak convergence to Browinan motion has been incorporated in the utilization of (one-sided) linear boundaries [and a more general spending function approach]. For rank based procedures, often, for not so large samples, permutation tools provide scope for good approximations. The spirit of UIP is inherent in such interim analysis too.

With the advent of genomics and bioinformatics, in general, clinical trials are also encountering some challeging tasks. Instead of the conventional symptomatic effect approach, there is a new emphasis on pharmacogenomics dealing with the drug responses and the detection of disease genes along with the gene-environment interaction. Recalling that there may be thousands of genes which in a polygenic mode may not have individually significant impact but a large number of them in synergy may have significant (joint) impact, clinical trials are charged with not only finding the genes associated (causally or statistically) with a specific (group of) disease(s) but also their pharmacokinetics and pharmacodynamics with specific drug development. Instead of clinical trials with human subjects it calls for additional refinements: microarray and proteomics studies in clinical trials setup at the molecular level with tissues or cells. While this subject matter is beyond the scope of the present study, at least, it could be emphasized that because of enormous cost in conducting such trials, multi-center trials are needed for pooling small information from the individual centers and also multiple end points typically arise in such composite studies. Typically, we encounter a matrix of statistics, individually from the centers and within each center, for the multiple end points. Although these centers may be treated as independent, the intra-center responses for the different end point are not. Confined to within center perspectives, typically, we have a vector valued stochastic process, and as before, we have a constrained environment. There-

fore, even if we are able to construct a martingale array (in a multi-dimensional setup), formulating CSI procedures in a proper manner could be a formidable task. Bessel process approximations for multi-dimensional stochastic processes in clinical trials have been studied in the literature (viz., Sen (1981, Chapter 11)). There is a challenging task of incorporating such distributional approximations in the formulation of statistical inference procedures for restrained environments. The prospects for multivariate CSI analysis, displayed in detail in Silvapulle and Sen (2004) need to be appraised further. It is our belief that UPI, because of its flexibility and amenity to more complex models, would be most suitable in this context too.

We conclude with some pertinent remarks on the role of UPI in meta analysis, as is currently adapted in multi-center clinical trials and genomic studies. Multi-center clinical trials, although, generally conducted under not so homogeneous environment (e.g., different geographical or demographic strata, age / culture differences), have a common objective of drawing statistical conclusions that pertain to a broader population. Consider in this vein, $C(\geq 2)$ centers, each one conducting a clinical trial with the common goal of comparing a new treatment with an existing one or a control or placebo. Since such centers pertain to patients with possibly different cultural, racial, demographic profiles, diet and physical exercise habits etc. and they may have somewhat different clinical norms too, the intra-center test statistics $\mathcal{L}_c,\ c = 1, \ldots, C$, used for CSI/RST, though could be statistically independent, might not be homogeneous enough to pull directly. This feature may thus create some impasses in combining these statistics values directly into a pooled one to enhance the statistical information. Meta analysis, based on observed significance levels (OSL) or p-values, is commonly advocated in this context. Recall that under the null hypothesis (which again can be interpreted as the intersection of all the null hypotheses), the p-values have the common uniform $(0, 1)$ distribution, providing more flexibility to adopt UIP in meta analysis. Under restricted alternatives, these OSL values are left-tilted (when appropriate UIT are used) in the sense that the probability density is positively skewed over $(0, 1)$ with high density at the lower tail and low at the upper. Let us denote the p-values by

$$P_c = P\{\mathcal{L}_c \geq \text{the observed value } |H_0\}, c = 1, \ldots, C. \tag{4.5}$$

The well-known Fisher's test is based on the statistic

$$F_n = \sum_{c=1}^{C} \{-2 \log P_c\}, \tag{4.6}$$

which, under the null hypothesis, has the central chi-square distribution with $2C$ degrees of freedom. This test has some desirable asymptotic properties. There are many other tests based on the OSL values. The well known step-

down procedure (Roy 1958) has also been adapted in this vein (cf. Subbaiah and Mudholkar 1980, Sen 1983), and they have been amended for CSI and RST as well (cf. Sen 1988). One technical drawback observed in this context is the insensitivity (to small to moderate departures from the null hypothesis) of such tests (including the Fisher's) when C is large, resulting in nonrobust and, to a certain extent, inefficient procedure. Thus, alternative approaches based on the OSL values have been explored more recently in the literature.

In the evolving field of bioinformatics and genomics, generally, we encounter an excessively high dimensional data set with inadequate sample size to induce the applicability of standard CSI or even conventional statistical inference tools. On top of that, in genomics, the OSL values to be combined (corresponding to different genes) may not be independent, creating another layer of difficulty with conventional meta analysis. This led to the development of multiple hypotheses testing in large dependent data models based on OSL values. This field is going through an evolution, and much remains to be accomplished. In this spectrum, the Simes (1986) theorem occupies a focal point. Let there be K null hypotheses (not necessarily independent) $H_{0k},\ k = 1, \ldots, K$ with respective alternatives (which possibly could be restricted or constrained as in clinical trials or microarray studies) $H_{1k},\ k = 1, \ldots, K$. We thus come across the same UIP scheme by letting H_0 as the intersection of all the component null hypotheses, and H_1 as the union of the component alternatives. Let $P_k, k = 1, \ldots, K$ be the OSL values associated with the hypotheses testing H_{0k} vs. H_{1k}, for $k = 1, \ldots K$. We denote the ordered values of these OSL values by $P_{K:1}, \cdots, P_{K:K}$. If the individual tests have continuous null distributions then the ties among the P_k (and hence, among their ordered values) can be neglected, in probability. Assuming independence of the P_k, Simes theorem states that

$$P\{P_{K:k} > k\alpha/K, \forall k = 1, \ldots, K | H_0\} = 1 - \alpha. \tag{4.7}$$

Interestingly enough, the Simes theorem is a restatement of the classical Ballot theorem, developed some twenty years before (cf. Karlin 1969). In any case, it is a nice illustration how the UIP is linked to the extraction of extra statistical information through ordered OSL values.

It did not take long time for applied mathematical statisticians to make good uses of the Simes-Ballot theorem in CSI and multiple hypothesis testing problems. The above results pertains to tests for an overall null hypothesis in the UIP setup. Among others, Hochberg (1988) considered a variant of the above result:

$$P\{P_{K:j} \geq \alpha/(K - j + 1), \forall j = 1, \ldots, K | H_0\} = 1 - \alpha, \tag{4.8}$$

and incorporated this result in a multiple testing framework. Benjamini and Hochberg (1995) introduced the concept of false discovery rate (FDR) in the

context of multiple hypothesis testing, and illustrated the role of the Simes-Ballot theorem in that context. The past ten years have witnessed a phenomenal growth of research literature in this subfield with applications to genomics and bioinformatics. The basic restraint in this respect is the assumption of independence of the $P_j, j = 1, \ldots, K$, and in bioinformatics, this is hardly the case. Sarkar (1998) and Sarkar and Chang (1997) incorporated the MTP_2 (multivariate total positivity of order 2) property to relax the assumption of independence to a certain extent. Sarkar (2000, 2002, 2004) has added much more to this development with special emphasis on controlling of FDR in some dependent cases. The literature is too large to cite adequately, but our primary emphasis here is to stress how UIP underlies some of these developments and to focus on further potential work.

Combining OSL values, in whatsoever manner, may generally involve some loss of information when the individual tests are sufficiently structured to have coherence that should be preserved in the meta analysis. We have seen earlier how guided by the UIP, progressive censoring in clinical trials provided more efficient and interpretable testing procedures. The classical Cochran-Mantel-Haenszel (CMH) procedure is a very notable example of this line of attack. In a comparatively more general multiparameter CSI setting, Sen (1999b) has emphasized the use of the CMH procedure in conjunction with the OSL values to induce greater flexibility. The field is far from being saturated with applicable research methodology. The basic assumption of independence or specific type of dependence is just a part of the limitations. A more burning question is the curse of dimensionality in CSI problems. Typically, there K is large and the sample size n is small, i.e., $K >> n$. In the context of clinical trials in genomics setups, Sen (2006) has appraised this problem with due emphasis on the UIP. Conventional test statistics (such as the classical LRT) have awkward distributional problems so that usual OSL values are hard to compute and implement in the contemplated CSI problems. Based on the Roy (1953) UIP but on some nonconventional statistics, it is shown that albeit there is some loss of statistical information due to the curse of dimensionality, there are suitable tests which can be implemented relatively easily in high-dimension low sample size environments. In CSI for clinical trials in the presence of genomics undercurrents, there is a tremendous scope for further developments along this line.

References

1. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B57,*, 289-300.

2. Chatterjee, S.K. and Sen, P.K. (1973). Nonparametric testing under progressive censoring. *Calcutta Statist. Assoc. Bull. 22*, 13-50.

3. Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc.B 34*, 187-220.

4. DeLong, D. M. (1981). Crossing probabilities for a square root boundary by a Bessel process. *Commun. Statist. A 10*, 2197-2213.

5. DeMets, D.L. and Lan, K.K.G. (1983). Discrete sequential boundaries for clinical trials. *Biometrika 70*, 659-663.

6. Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*,800-802.

7. Karlin, S. (1969). *A first Course in Stochastic Processes*, Academic Press, New York.

8. Roy, J. (1958). Step-down procedures in multivariate analysis. *Ann. Math. Statist. 29*, 1177-1188.

9. Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.24*, 220-238.

10. Sarkar, S.K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture.*Ann. Statist.26*, 494-504.

11. Sarkar, S.K. (2000). A note on the monotonicity of the critical values of a step-up test. *J. Statist. Plann. Infer. 87*, 241-249.

12. Sarkar, S.K. (2002). Some results on false discovery rate in multiple testing procedures. *Ann. Statist. 30*, 239-257.

13. Sarkar, S.K. (2004). FDR-controlling stepwise procedures and their false negativesc rates. *J. Statist. Plann. Infer. 125*, 119-137.

14. Sarkar, S.K. and Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc. 92*, 1601-1608.

15. Sen, P.K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference* John Wiley, New York.

16. Sen, P.K. (1983). A Fisherian detour of the step-down procedure. In *Contributions to Statistics: Essays in honour of Norman L. Johnson* North Holland, Amsterdam, pp. 367-377.

17. Sen, P.K. (1988). Combination of statistical tests for multivariate hypotheses against restricted alternatives. In *Advances in Multivariate Statistical Analysis* (eds. S.Dasgupta and J.K. Ghosh), Ind. Statist. Inst. pp. 377-402.

18. Sen, P. K. (1999a). Multiple comparisons in interim analysis. *J. Statist. Plann. Infer. 82*, 5-23.

19. Sen, P.K. (1999b). Some remarks on the Stein-type multiple tests of significance. *J. Statist. Plann. Infer. 82*, 139-145.

20. Sen, P. K. (2001). Survival analysis: Parametrics to semiparametrics to pharmacogenomics. *Brazilian J. Probab. Statist. 15*, 201 - 220.

21. Sen, P. K. (2006). Robust Statistical inference for high-dimension low sample size problems with applications to genomics. *Austrian J. Statist.* in press.

22. Silvapulle, M.J. and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order and Shape Restrictions* John Wiley, New York.

23. Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751-754.

24. Subbaiah, P. and Mudholkar, G. S. (1980). Testing significance of a mean vector - a possible alternative to Hotelling T^2. *Ann. Inst. Statist. Math. 32*, 43-52.

25. Temple, R. and Ellenberg, S.S. (2000). Placebo-controlled trials and active-controlled trials in the evaluation of new treatments, I: Ethical and scientific issues. *Ann. Inter. Med. 133*, 455-463.

26. Tsai, M.-T. and Sen, P.K. (2005). Asymptotically optimal tests for parametric functions against ordered functional alternatives. *J. Multivar. Anal. 95*, 37-49.

Part II

Invited Papers

5

Parametric Accelerated Life Model in Survival Analysis

V. Bagdonavicius[†], L. Clerjaud[‡], and M. Nikulin[‡]

[†] *Vilnius University*
[‡] *Bordeaux II University*

Abstract: We consider the application of the parametric Accelerated Life model in the Accelerated Life Testing, and we suppose that the failures times have Generalized Weibull distributions. The hazard rates can be IFR or DFR (Weibull, G.W), $\cap -$ *shaped* (log-normal, Generalized Weibull) or $\cup -$ *shaped* (Generalized Weibull only). We propose estimation procedure to the AFT-GW model in case of constant stress or step-stress.

5.1 The Accelerated Failure Time Model

5.1.1 Definitions and notations

Let $x(\cdot)$ be a *m-dimensional time-dependent stress* : $x(\cdot) = (x_0(\cdot), \ldots, x_m(\cdot))^T$, where $x_1(\cdot), \ldots, x_m(\cdot)$ are the *univariate time-dependent stresses*. Let E be a set of all admissible (possible) m-dimensional stresses:

$$E = \left\{ x(\cdot) : x(\cdot) = (x_1(\cdot), \ldots, x_m(\cdot))^T ; x : [0, +\infty[\longrightarrow R^m \right\}$$

The application of accelerated stresses shortens the life durations of units. Let x_0 be the usual (standard or normal) stress: $x_0 \in E_1 \subset E$, E_1 is a set of all constant-in-time stresses. Also let us suppose that the lifetime $T_{x(\cdot)}$ under stress $x(\cdot)$ is a positive random variable, and $S_{x(\cdot)}$ be the survival function of $T_{x(\cdot)}$:

$$S_{x(\cdot)}(t) = P\left\{T_{x(\cdot)} \geq t\right\} \tag{1}$$

The AFT model is true on E , if there exists a function $r : E \longrightarrow R_+$ such that

$$S_{x(\cdot)}(t) = S_0\left(\int_0^t r[x(\tau)]d\tau\right), x(\cdot) \in E \tag{2}$$

where S_0 is the baseline survival function. It could be possible that $S_0 = S_{x_0}$. In parametric models S_0 belongs to a *parametric family*, and $r[x(\cdot)]$ is

parameterized. If $x(\tau) = x \in E_1$ is constant in time, then we get from the equation (2), the model by different way.

$$S_x(t) = S_0(r(x)t), x \in E_1 \tag{3}$$

If the stress x is one-dimensional (scalar) and constant, r is often parameterized as

$$r(x) = e^{-\beta_0 - \beta_1 \varphi(x)} \tag{4}$$

where φ is a given function of x. The most applied are 3 famous models:

- *log-linear model* , where $r(x) = e^{-\beta_0 - \beta_1 x}$, $\varphi(x) = x$, (4a)
- *power-rule model*, where $r(x) = e^{-\beta_0 - \beta_1 \ln x}$, $\varphi(x) = \ln x$, $x > 0$, (4b)
- *Arrhenius model*, where $r(x) = e^{-\beta_0 - \beta_1 / x}$, $\varphi(x) = 1/x$, x is a scalar. (4c)

5.1.2 Plan of experiments in parametric AFT model

Let us consider the following *first plan of experiments* when the items are tested under the accelerated m-dimensional constant stresses $x_1, \dots x_k$. The *usual stress* x_0 is *not used* during the experiments. Let k be the number of observed groups of units. n_i units are tested under stress $x_i > x_0$, $(i = 1, \dots, k)$. We consider here 3 cases where the baseline survival function S_0, belongs to the family of *Generalized Weibull* distribution, such as:

$$S_0(t) = exp\left\{1 - (1 + (t/\theta)^\nu)^\gamma\right\}, t > 0, \nu > 0, (\textit{Generalized Weibull}). \tag{5}$$

θ parameter may be included in the parameter β_0 in the AFT model. These survival function S_0 can be respectively expressed as:

$$S_0(t) = exp\left\{1 - (1 + t^\nu)^\gamma\right\}, t > 0, \nu > 0, (\textit{Generalized Weibull}). \tag{6}$$

In the case of $\gamma = 1$, we have the Weibull distribution.

Let us consider the *second plan* of experiments when the items are tested under the constant step-stresses, where n units are on test at an initial low stress. And if it does not fail in a predetermined time t_1 is increased and so on. Thus all units are tested under the step-stress

$$x(\tau) = \begin{cases} x_1, 0 \le \tau < t_1, \\ x_2, t_1 \le \tau < t_2, \\ \dots, \dots \\ x_k, t_{k-1} \le \tau < t_k, \end{cases} \tag{7}$$

where x_j are m-dimensional constant stresses,$t_0 = 0$,$t_k = +\infty$. Thus , for step-stresses, the survival function can be written as

$$S_{x(\cdot)}(t) = S_{x_i}\left\{t - t_{i-1} + \frac{1}{r(x_i)}\sum_{j=1}^{i-1} r(x_j)(t_j - t_{j-1})\right\}, \tag{8}$$

where $i = 1, 2, \ldots, k$ and $t \in [t_{i-1}, t_i[$. Therefore the survival function under stress $x(\tau)$ is

$$S_{x(.)}(t) = S_0\left\{\mathbf{1}\{i>1\}\sum_{j=1}^{i-1} e^{-\beta^T x_j}(t_j - t_{j-1}) + e^{-\beta^T x_i}(t - t_{i-1})\right\},$$

where x_j may be $\varphi(x_j)$.

In the case of Generalized Weibull distribution, if $0 < \nu < 1$ and $1/\gamma < \nu$, the curve of the hazard function has a $\cup -$ *shape*. If $1/\gamma > \nu > 1$, the curve of hazard function has a $\cap -$ *shape*. If $0 < \nu < 1$ and $\nu < 1/\gamma$ then the hazard function decreases from $+\infty$ to 0 (DFR). If $\nu > 1$ and $\nu > 1/\gamma$ then hazard function increase to $+\infty$ (IFR).To estimate this survival function under the stress x and its two confidence limits, we have to estimate the parameter β and σ(and possibly γ).

5.1.3 Parameter estimation: Generalized Weibull distribution

Let us consider the first plan of experiments when the units are tested under accelerated constant stresses. Let t_i be the maximal experiment duration for the ith group. n_i units are tested under accelerated stress x_i $(i = 1, \ldots, k)$ Let $\beta = (\beta_0, \ldots, \beta_m)^T$ be the regression parameter. The lifetime of the jth unit from ith group is the variable T_{ij}. Set $X_{ij} = T_{ij} \wedge t_i$ and $\delta_{ij} = \mathbf{1}\{T_{ij} < t_i\}$.

In case of various stress, the likelihood function is:

$$L(\beta,\nu,\gamma) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\left\{\nu\gamma\left(1+\left(f_i(X_{ij},\beta,\gamma)\right)^{\nu}\right)^{\gamma-1} e^{-\beta^T x^{(i)}(X_{ij})}\left(f_i(X_{ij},\beta,\gamma)\right)^{\nu-1}\right\}^{\delta_{ij}} \times \tag{9}$$

$$exp\left\{1-\left(1+\left(f_i(X_{ij},\beta,\gamma)\right)^{\nu}\right)^{\gamma}\right\}$$

where

$$f_i(X_{ij},\beta,\gamma) = \int_0^{X_{ij}} e^{-\beta^T x^{(i)}(u)}du$$

In case of constant stress, we get:

$$L(\beta,\nu,\gamma) = \prod_{i=1}^{k}\prod_{j=1}^{n_i} \nu\gamma e^{-\nu\beta^T x^{(i)}} X_{ij}^{\nu-1}\left(1+e^{-\beta^T x^{(i)}} X_{ij}^{\nu}\right)^{\gamma-1}\right)^{\delta_{ij}} exp\left\{1-\left(1+e^{-\beta^T x^{(i)}} X_{ij}^{\nu}\right)^{\gamma}\right\} \tag{10}$$

where $x^{(i)} = (x_{i0}, \ldots, x_{im})$, $x_{i0} = 1$ and $\nu = 1/\sigma$.

By derivations of this log-likelihood $\ln L(\beta,\nu,\gamma)$, we get :

$$U_l(\beta,\nu,\gamma) = \frac{\partial \ln L(\beta,\nu,\gamma)}{\partial\beta_l} = \nu\sum_{i=1}^{k} x_{il}\sum_{j=1}^{n_i}\left(\gamma\omega_{ij}(\beta,\nu,\gamma) - \delta_{ij}u_{ij}(\beta,\nu,\gamma)\right)\ l = 0,\ldots,m; \tag{11a}$$

$$U_{m+1}(\beta,\nu,\gamma)=\frac{\partial \ln L(\beta,\nu,\gamma)}{\partial \nu}=\frac{D}{\nu}-\frac{1}{\nu}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)-\delta_{ij}u_{ij}\left(\beta,\nu,\gamma\right)\right)\ln h_{ij}(\beta,\nu,\gamma) \tag{11b}$$

$$U_{m+2}(\beta,\nu,\gamma)=\frac{\partial \ln L(\beta,\nu,\gamma)}{\partial \gamma}=\frac{D}{\nu}-\frac{1}{\nu}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\left(1+h_{ij}\left(\beta,\nu\right)\right)^{\gamma}-\delta_{ij}\right)\ln\left(1+h_{ij}(\beta,\nu)\right), \tag{11c}$$

where

$$D=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\delta_{ij}$$

and

$$h_{ij}(\beta,\nu)=\left(e^{-\beta^T x^{(i)}}X_{ij}\right)^{\nu},\ \omega_{ij}\left(\beta,\nu,\gamma\right)=\left(1+h_{ij}(\beta,\nu)\right)^{\gamma-1},\ u_{ij}(\beta,\nu,\gamma)=1+(\gamma-1)\frac{h_{ij}(\beta,\nu)}{1+h_{ij}(\beta,\nu)}.$$

The Fisher information $I(\beta,\nu,\gamma)=\left(I_{ls}\left(\beta,\nu,\gamma\right)\right)_{(m+3)\times(m+3)}$ is a matrix with the following components :

$$I_{ls}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2 \ln L\left(\beta,\nu,\gamma\right)}{\partial\beta_l\partial\beta_s}=\nu\sum_{i=1}^{k}x_{il}x_{is}\times \tag{12a}$$

$$\sum_{j=1}^{n_i}\ \frac{\nu\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+h_{ij}\left(\beta,\nu\right)\right)-\delta_{ij}\left(u_{ij}\left(\beta,\nu,\gamma\right)-1\right)}{1+h_{ij}\left(\beta,\nu\right)}\ ,\ (l,s=0,\ldots,m)$$

$$I_{l,m+1}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2 \ln L\left(\beta,\nu,\gamma\right)}{\partial\beta_l\partial\nu}=-\frac{1}{\nu}U_l\left(\beta,\nu,\gamma\right)-\sum_{i=1}^{k}x_{il}\sum_{j=1}^{n_i}\ \frac{\ln h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\ \times \tag{12b}$$

$$\left\{\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+\gamma h_{ij}\left(\beta,\nu\right)\right)-\delta_{ij}\left(u_{ij}\left(\beta,\nu,\gamma\right)-1\right)\right\},\ (l=0,\ldots,m)$$

$$I_{l,m+2}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2 \ln L\left(\beta,\nu,\gamma\right)}{\partial\beta_l\partial\gamma}=-\frac{1}{\nu}\sum_{i=1}^{k}x_{il}\times \tag{12c}$$

$$\sum_{j=1}^{n_i}\ \omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+\gamma\ln\left(1+h_{ij}\left(\beta,\nu\right)\right)\right)-\delta_{ij}\frac{h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\ \ ,\ (l=0,\ldots,m),$$

$$I_{m+1,m+1}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2 \ln L\left(\beta,\nu,\gamma\right)}{\partial\nu^2}=\frac{1}{\nu}U_{m+1}\left(\beta,\nu,\gamma\right)+\frac{1}{\nu^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ \frac{\ln^2 h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\ \times$$

$$\left\{\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+\gamma h_{ij}\left(\beta,\nu\right)\right)-\delta_{ij}\left(u_{ij}\left(\beta,\nu,\gamma\right)-1\right)\right\}+ \tag{12d}$$

$$+\frac{1}{\nu^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln h_{ij}\left(\beta,\nu,\gamma\right)\left\{\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)-\delta_{ij}u_{ij}\left(\beta,\nu,\gamma\right)\right\},$$

$$I_{m+1,m+2}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2 \ln L\left(\beta,\nu,\gamma\right)}{\partial\nu\partial\gamma}=\frac{1}{\nu}\times\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{h_{ij}\left(\beta,\nu\right)\ln h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\times \tag{12e}$$

$$\{(1+h_{ij}(\beta,\nu))^{\gamma}+\gamma(1+h_{ij}(\beta,\nu))^{\gamma}+\ln(1+h_{ij}(\beta,\nu))-\delta_{ij}\}$$

$$I_{m+2,m+2}(\beta,\nu,\gamma)=-\frac{\partial^2 \ln L(\beta,\nu,\gamma)}{\partial\gamma^2}=\frac{D}{\gamma^2}+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(1+h_{ij}(\beta,\nu,\gamma))^{\gamma}\ln^2(1+h_{ij}(\beta,\nu,\gamma)) \quad (12f)$$

Then, the estimated survival function under usual stress $x^{(0)}$ is

$$\hat{S}_{x^{(0)}}(t)=exp\left\{1-\left(1+\left(e^{-\hat{\beta}^T x^{(0)}}t\right)^{\hat{\nu}}\right)^{\hat{\gamma}}\right\} \quad (13)$$

where $x^{(0)}\in E_0$ is the usual stress (standard or normal).

Under the stress x, the $(1-\alpha)$ approximate confidence limit for $Q_x(t)=\ln(S_x(t)/(1-S_x(t)))$ and $S_x(t)$ is respectively

$$\hat{Q}_x(t)\pm\hat{\sigma}_{Q_x}w_{1-\alpha/2} \quad (14)$$

and

$$\left(1+\frac{1-\hat{S}_x(t)}{\hat{S}_x(t)}exp\left\{\pm\hat{\sigma}_{Q_x}w_{1-\alpha/2}\right\}\right) \quad (15)$$

where

$$\hat{\sigma}_{Q_x}=\frac{1}{1-\hat{S}_x{}^2}\sum_{l=0}^{m+2}\sum_{s=0}^{m+2}a_l\ \ \hat{\beta},\hat{\nu},\hat{\gamma}\ \ I^{ls}\ \ \hat{\beta},\hat{\nu},\hat{\gamma}\ \ a_s^T\ \ \hat{\beta},\hat{\nu},\hat{\gamma}$$

$$a_l\ \ \hat{\beta},\hat{\nu},\hat{\gamma}\ \ =-\hat{\nu}x_l a_{m+1}\ \ \hat{\beta},\hat{\nu},\hat{\gamma}\ \ /\ \ \ln t-\hat{\beta}^T x\ \ ,l=0,\ldots,m$$

$$a_{m+1}\ \ \hat{\beta},\hat{\nu},\hat{\gamma}\ \ =-\hat{\gamma}\ \ e^{-\hat{\beta}^T x}t^{\ \hat{\nu}}\ \ 1+\ \ e^{-\hat{\beta}^T x}t^{\ \hat{\nu}}\ \ ^{\hat{\gamma}-1}\ \ \ln t-\hat{\beta}^T x\ \ ,$$

$$a_{m+2}\ \ \hat{\beta},\hat{\nu},\hat{\gamma}\ \ =-\ \ 1-\ln\hat{S}_x(t)\ \ \ln\ \ 1+\ \ e^{-\hat{\beta}^T x}t^{\ \hat{\nu}}$$

The $(1-\alpha)$ approximate confidence limits for $Q_x(t)=\ln(S_x(t)/(S_x(t)))$ and $S_x(t)$ are respectively the same with $x=x^{(0)}$.

5.2 Results

5.2.1 Generalized Weibull, Weibull or Log-normal failure time distributions in the case of $\cup-$*shaped* hazard rate function

The aim of this example is to show which distribution is better in fitting the survival function when the hazard rate function has a $\cup-$*shape*. We simulated data from Generalized Weibull distribution with parameters $\gamma=6.5$ and $\nu=0.7$, $\beta_0=9$ and $\beta_1=-0.8$. The values of one-dimensional stress x are : $x_1=2<x_2=4<x_3=6<x_4=10$.These observations are not censored and the power-rule model is used. Let $x_0<x_1$ be the usual stress. There are $n_1=3300$

observations under the stress x_1 , $n_2 = 2900$ observations under the stress x_2, $n_3 = 2000$ observations under the stress x_3, $n_4 = 1300$ observations under the stress x_4. The estimator s of the parameters and the log-likelihood are given in Table 5.1, after simulating 600 times by using Monte-Carlo method. The

Table 5.1: Estimators of parameters from data with $\cup$ – *shaped* hazard rate function.

Distributions	**Weibull**	**Generalized Weibull**	**Log-normal**
Log Likelihood	-54567.81	-54313.08	-56092.1
$\hat{\beta}_0$	5.7909955	8.9048171	5.1076953
$\hat{\beta}_1$	-0.8005198	-0.7805359	-0.797745
$\hat{\nu}$	0.8968277	0.6998924	0.6313976
$\hat{\gamma}$	None	6.2579115	None
$\hat{m}\left(x^{(0)}\right)$	145.3675	241.1373	171.5597

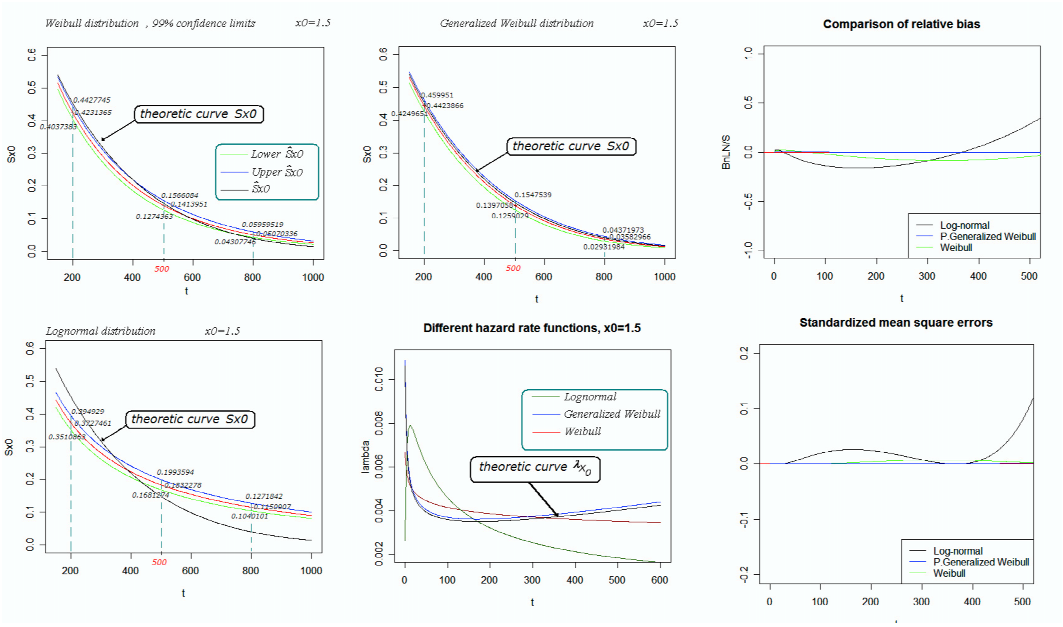


Figure 5.1: Different curves of S_{x_0} and λ_{x_0},different curves of relative bias following the failure time distribution.

estimators of the survival function under usual stress with two 99% confidence limits are given in Figure 5.1. In case of Generalized Weibull distribution, the two 99% confidence limits a little narrower than two confidence limits in case of Weibull and Log-normal distributions. In several moments t, in the case of the AFT-Weibull and AFT-log-normal models, the theoretic curve of S_x is out of the two 99% confidence limits, whereas it is not the case for AFT-Generalized

Weibul model.

The bias, between the real value of survival function $S_x(t, \theta)$ and the mean of estimated $\hat{S}_x(t)$ under the stress x, is shown as in Figure 5.1. From the Figure 5.1, the bias is lower in case of Generalized Weibull distribution than in cases of Weibull and Log-normal distributions. We see that the Generalized Weibull distribution is quite better than Weibull and Log-normal distributions, because this distribution is the only one which fits better this survival function when the hazard rate function has a $\cup -$ *shape*. If we use the likelihood ratio test, like

$$Z = -2\left(\ln LL_W\left(\hat{\mu}\right) - \ln LL_{GW}\left(\hat{\theta}\right)\right)$$

we see that Z takes the high values, and its significance is $0 < 0.05$. For 600 simulations, there are 600 rejections of hypothesis that there are no difference between Generalized Weibull and Weibull distributions, the power test is 1. The bias is quite smaller than other cases, if we suppose that this hazard rate function is $\cup -$ *shaped*. And the theoretic curve is not out of the two 99% confidence limits. Therefore, the Generalized Weibull fits much better than Weibull and Lognormal distribution.

5.2.2 Generalized Weibull distribution in case of step-stress

We are interested in the case of step-stress, we can see how the survival function could behave, in Figure 5.2. For the size of 8000 items, the values of estimated

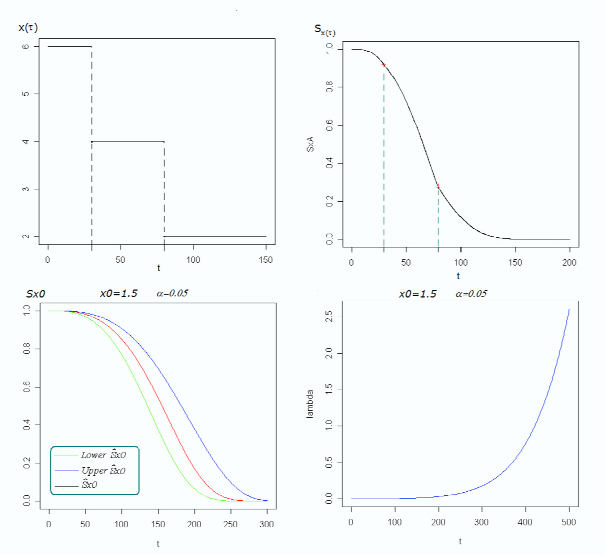


Figure 5.2: Survival functions with step-stress, IFR hazard rate function.

parameters $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\nu}$ and $\hat{\gamma}$ are respectively 5.7888206, -0.7337947, 3.0522533 and 2.3047733. The value of log-likelihood is -37341.98. We see that the two 95% confidence limits of S_{x_0} are wider than the case of constant stress for the same size of sample. The width of these two confidence limits are due to the fact that the step-stress makes the estimation of survival functions more uncertain.

References

1. Nelson, W. (2004). Accelerated Testing, Statistical Models, Test Plans, and Data Analysis. John Wiley & Sons, New Jersey. 601p.
2. Bagdonavicius, V. Clerjaud, L, Nikulin, M.S. (2006) Power Generalized Weibull in Accelerated Life Testing, Preprint N°0602, Université Victor Segalen Bordeaux 2, France
3. Bagdonavicius, V. Nikulin, M. (2002) Accelerated Life Models, Modeling and Statistical Analysis. Chapman & Hall/CRC. 334p.
4. Meeker, W.Q. Escobar, L. (1998). Statistical Method for Reliability Data. J.Wiley, New York.
5. Kishore, Sundara Rajan. Accelerated Testing-An EE perspective, Preprint, August 2002.
6. Wilson, S.P. (2000) Failure Model Indexed by Time and usage. In : Recent Advances in Reliability Theory, Methodology, Practice and Inference, (Eds. N.Limnios and ML.Nikulin) Boston : Birkhauser, 213-218.
7. Kahle W, Wendt, H.(2000) Statistical Analysis of Damage Process. In: Recent Advances in Reliability Theory, Methodology, Practice and Inference, (Eds. N.Limnios and M.Nikulin) Boston : Birkhauser, 199-212.
8. Viertl, R. (1988). Statistical Methods in Accelerated Life Testing. Vandenhoeck & Ruprecht, Gottingen.
9. Xiong, C. and MING, J. "Analysis of Grouped and Censored Data from Step-Stress Life Test", IEEE Trans. On Rel., Vol. 53, No. 1, p.22-28, March 2004.

6

Exact Inference and Optimal Censoring Scheme for a Step-Stress Model Under Progressive Type-II Censoring

N. Balakrishnan

Department of Mathematics and Statistics,
McMaster University,
Hamilton, Ontario, Canada L8S 4K1

Abstract: In reliability and life-testing experiments, the researcher is often interested in the effects of extreme or varying stress factors such as temperature, voltage, and load on the lifetimes of experimental units. Step-stress test, which is a special case of accelerated life-tests, allows the experimenter to increase the stress levels at fixed times during the experiment in order to obtain information on the parameters of the life distribution more quickly than under normal operating conditions.

In this talk, I will consider a simple step-stress model under the exponential distribution when the available data are progressively Type-II censored, and obtain the maximum likelihood estimators (MLEs) of the parameters assuming a cumulative exposure model with lifetimes being exponentially distributed. I will derive the exact distributions of the MLEs of parameters through the use of conditional moment generating functions and discuss the construction of confidence intervals for the parameters using these exact distributions, asymptotic distributions of the MLEs, and the parametric bootstrap methods, and then evaluate the performance of all these confidence intervals based on an extensive Monte Carlo simulation study. Next, I will discuss the determination of optimal progressive censoring schemes as well as the optimal time for change of the stress level based on the simple step-stress model. Finally, I will present a few examples to illustrate all the methods of inference developed here.

Keywords and phrases: Accelerated testing; Bootstrap method; Conditional moment generating function; Coverage probability; Cumulative exposure

model; Exponential distribution; Maximum likelihood estimation; Optimal censoring scheme; Order statistics; Step-stress models; Tail probability; Progressive Type-II censoring

7

Non-Periodic Inspections To Guarantee A Prescribed Level Of Reliability

C.T. Barker and M.J. Newby

The City University
School of Engineering and Mathematical Sciences
Northampton Square, London EC1V 0HB, England
Emails: c.t.barker@city.ac.uk and m.j.newby@city.ac.uk

Abstract: A cost optimal non-periodic inspection policy is derived for complex multi-component systems. The model takes into consideration the degradation of all the components in the system with the use of a Bessel process with drift. The inspection times are determined by a deterministic function of the system state. The non-periodic policy is developed by evaluating the expected lifetime costs and the optimal policy by an optimal choice of inspection function. The model thus gives a guaranteed level of reliability throughout the life of the project.

Keywords and phrases: Wiener process; Bessel process; regenerative process; renewal-reward

7.1 Introduction

The aim of the paper is to derive a cost-optimal inspection and maintenance policy for a multi-component system whose state of deterioration is modelled with the use of a Markov stochastic process. Each component in the system undergoes a deterioration described by a Wiener process. The proposed model takes into account the different deterioration processes by considering a multivariate state description $\mathbf{W}_t$. The performance measure R_t of the system is a functional on the underlying process and is not monotone. Because we now wish to ensure a minimum level of reliability is maintained, we set the critical threshold at an acceptable level and examine the probability that the system will never return to this level after crossing it. When this occurs, the system is aging in such a way that it needs to be repaired or replaced.

7.2 The Model

In this section, we define the considered processes used in our model. We also explain how the maintenance actions and the non-periodic inspections are modelled.

7.2.1 The considered processes

The state of each component is modelled with the use of a Wiener process

$$W_t^{(i)} = \mu_i t + \sigma B_t^{(i)},\ W_0^{(i)} = 0,\ \forall i \in \{1, \ldots, N\} \tag{7.1}$$

where $B_t^{(i)}$ are independent Brownian motions.
The system's state of deterioration is described by the corresponding multivariate Wiener process:

$$\mathbf{W}_t = \left(W_t^{(1)}, W_t^{(2)}, \ldots, W_t^{(N)}\right) \tag{7.2}$$

When the system is inspected a performance measure is calculated. This performance measure is a functional on the underlying process, its Euclidean norm:

$$R_t = \quad \|\mathbf{W}_t\|_2 = \sqrt{\sum_{i=1}^{N} (W_t^{(i)})^2} \tag{7.3}$$

Thus, R_t is $Bes_0(\nu, \mu)$: the Bessel process starting at the origin with parameter ν, drift μ where:

$$\nu = \frac{1}{2}N - 1,\ \mu = \sqrt{\sum_{i=1}^{N} \mu_i^2} \tag{7.4}$$

We refer the reader to Pitman J.W. and Yor M. (1981) and Revuz D. and Yor M. (1991) for further details on the Bessel process.

The non-monotonicity of the performance measure R_t is handled by defining a critical threshold ξ, which determines the response to an inspection. Because we now wish to ensure a minimum level of reliability is maintained, we set the critical threshold at an acceptable level and examine the probability that the system will never return to this level after crossing it. For this we consider the following process:

$$H_\xi^0 = \sup_{t \in R^+} \{t : R_t > \xi | R_0 = 0\} \tag{7.5}$$

The probability density functions for both R_t and H_ξ^0 are known and given in Pitman J.W. and Yor M. (1981) and Barker C. and Newby M. (2006)

7.2.2 Non-periodic inspections and maintenance actions

The efficiency of the proposed policy entirely depends on the inspection times and the type of maintenance on the system. Maintenance on the system is modelled with the help of a maintenance function d. It is a decreasing bijective function of the state of the system at inspection times only. Correspondingly, we define a maintenance cost function C_R depending on the state of the process too. In the numerical example of section 7.4, the considered functions d and C_R are:

$$d(y) = \quad y \quad , \text{ if } y < \frac{\xi}{2} \tag{7.6}$$

$$= \quad k * y \quad , \text{ if } y > \frac{\xi}{2} \tag{7.7}$$

for both $k = 0.9$ and $k = 0.1$,

$$C_R(y) = \quad 0 \quad , \text{ if } X_\tau < \frac{\xi}{2} \tag{7.8}$$

$$= \quad 100 \quad , \text{ if } X_\tau \geq \frac{\xi}{2} \tag{7.9}$$

This paper considers a non-periodic inspection policy. The reason for this is that it is more general and often more useful than the periodic policy, since it generally results in policies with lower costs. This is done by considering the inspection scheduling function m first introduced by Grall *et al.* (2002). It is a decreasing function of the state of the process $d\left(X_t\right)$ after a maintenance action and determines the amount of time until the next inspection time. If we let T_i denote the times at which the system is inspected, we have:

$$T_{i+1} = T_i + m\left(X_{d\left(X_{T_i}\right)}\right) \tag{7.10}$$

The approach is to optimize the total expected cost with respect to the inspection scheduling function. The inspection functions form a two parameter family and these two parameters are allowed to vary to locate the optimum values. The function can be written $m(x \,|\, a,\, b)$ leading to a total cost function $C(a, b)$ which is optimized with respect to a and b. Different forms of inspection functions were considered with different convexity properties. Numerical examples shown in the present article consider the following inspection scheduling function:

$$m\left(x\right) = \quad -\left(\tfrac{\sqrt{a-1}}{b}x\right)^2 + a \quad , \text{ if } 0 \leq x \leq b \tag{7.11}$$

$$= \qquad 1 \qquad\quad , \text{ if } x \geq b \tag{7.12}$$

7.3 Expected Total Cost

This section gives the expression for the expected total cost. It is derived by considering the different scenarios at inspection times:

- $\mathbf{1}_{\{H^0_{\xi-x}>m(x)\}} = 1$
- $\mathbf{1}_{\{H^0_{\xi-x}>m(x)\}} = 0$

7.3.1 Expression of the expected total cost

We give the expression for the expected total cost without details, these are similar to the ones given in Barker C. and Newby M. (2006):

$$v^{(0)}_{\xi-x} = Q(x) + \lambda(x)\, v^{(0)}_{\xi} + \int_0^{d^{-1}(\xi)} v^{(0)}_{\xi-d(y)} K(x,y)\, dy \tag{7.13}$$

where:

$$\lambda(x) = \int_0^{m(x)} h^0_{\xi-x}(t)\, dt \tag{7.14}$$

$$Q(x) = (1-\lambda(x)) \left(C_i + \int_0^{+\infty} C_R(y)\, f^0_{m(x)}(y)\, dy \right) + C_f \lambda(x) \tag{7.15}$$

$$K(x,y) = \left(1 - \int_0^{m(x)} h^0_{\xi-x}(t)\, dt \right) f^0_{m(x)}(y) \tag{7.16}$$

7.3.2 Obtaining the solutions

The equation Eq. 7.13 is solved numerically. The method used here is similar to one given in Press W. H. *et al.* (1992). First, note that at $t = 0$ the system is new. Under this condition, we rewrite equation Eq. 7.13 as follows:

$$v^{(0)}_{\xi-x} = Q(x) + \lambda(x)\, v^{(0)}_{\xi-x} + \int_0^{d^{-1}(\xi)} v^{(0)}_{\xi-d(y)} K(x,y)\, dy \tag{7.17}$$

Rewriting 7.13 as 7.17 does not affect the solution to the equation and will allow the required solution to be obtained by a homotopy argument based on ξ. The Nystrom routine with the N-point Gauss-Legendre rule at the points $y_j, j \in \{1,\ldots,N\}$ is applied to 7.17, we get

$$\{1-\lambda(x)\} v^{(0)}_{\xi-x} = Q(x) + \sum_{j=1}^{N} v^{(0)}_{\xi-d(y_j)} K(x,y_j)\, w_j \tag{7.18}$$

We then evaluate the above at the following appropriate points $x_i = d(y_j)$, Eq. 7.18 can thus be rewritten in the following matrix form:

$$\left(\mathbf{D} - \widetilde{\mathbf{K}}\right) \mathbf{v} = \mathbf{Q} \tag{7.19}$$

where:

$$\mathbf{D}_{i,j} = (1 - \lambda(x_i))\, \mathbf{1}_{\{i=j\}},\ \widetilde{\mathbf{K}}_{i,j} = K(x_i, y_j)\, w_j,\ \mathbf{Q}_i = Q(x_i) \tag{7.20}$$

Having obtained the solution at the quadrature points by solving inversion of the matrix $\mathbf{D} - \widetilde{\mathbf{K}}$, we get the solution at any other quadrature point x by simply using equation Eq. 7.18 as an interpolatory formula, hence at the desired quadrature point $x_i = 0$.

7.4 Numerical Results and Comments

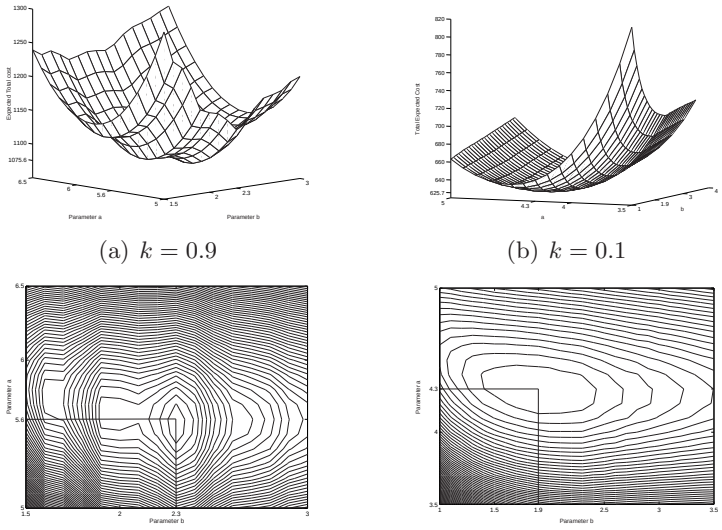
This section presents results from numerical experiments. The values of the parameters for the process used to model the degradation of the system and the different costs used were chosen arbitrarily to show some important features of the inspection policy. The initial value for the critical threshold is $\xi = 5$, the Bessel process considered is $Bes_0(0.5, 1)$ and the value for the cost of inspection and the cost of failure are $C_i = 50$ and $C_f = 200$. The numerical results for the case of small maintenance on the system ($k = 0.9$) and the case of a large amount of maintenance ($k = 0.1$) are shown in Fig. 7.1.

We first note that the surfaces and the contours clearly show the presence of an optimal policy. Moreover, in the case $k = 0.9$, we note the presence of a global minimum and also multiple local minima. This gives the decision maker some kind of choice in the inspection policy to consider.

Inspection policies		k=0.9	k=0.1
m	a^*	5.6	4.3
	b^*	2.3	1.9
	v^*	1075.6	625.6727

Table 7.1: Optimal values of the parameters a and b

The optimal values a^*, b^* and v^* for a, b and v_ξ^0 respectively, in the two different scenarios, can be summarized in Table 7.1. We note that the optimal cost are smaller for $k = 0.1$ than for $k = 0.9$. This makes sense, since in both cases the same values for the costs were considered: as the case $k = 0.1$ corresponds to bigger amounts of repair, the system will tend to deteriorate slower and therefore will require less maintenance resulting in a smaller total cost.



(a) $k = 0.9$ (b) $k = 0.1$

(c) $k = 0.9$ (d) $k = 0.1$

Figure 7.1: Surfaces and contour representations for expected total costs with $k = 0.9$ and $k = 0.1$

References

1. Barker C.T. and Newby M.J. (2006). Inspection and Maintenance Planning for Complex Multi-Component Systems, to appear in *Proceedings of ALT'2006, ISTIA*, Angers, France

2. Revuz D. and Yor M. (1991). *Continuous Martingale and Brownian Motion*, Springer Verlag.

3. Pitman J. W. and Yor M. (1981). Bessel Process and Infinitely Divisible Laws, Stochastic Integrals, *Lecture Notes in Mathematics, Springer Berlin.*

4. Press W. H. , Teukolsky S. A. ,Vetterling W. T. and Flannery B. P. (1992). *Numerical recipes in C, 2nd Edition*, Cambridge Univerity Press.

5. Grall A. , Dieulle L., Berenger C. and Roussignol M. (2002). Continuous-Time Predictive-Maintenance Scheduling for a Deteriorating System, *IEE Transaction on Reliability*, Vol. 51, No. 2.

8

A Bayesian Chi-squared Test for Censored Data

Calle, M.L. and Gómez, G.

Systems Biology Department, Universitat de Vic
Statistics and Operational Research Department, Universitat Politècnica de Catalunya

Abstract: The analysis of censored data has been mainly approached through nonparametric or semiparametric methods. One of the reasons of the widely use of these methods as opposed to classical parametric approaches relies in the difficulty of checking the validity of a parametric model when data are censored. In this work we propose a Bayesian chi-squared test of goodness-of-fit for censored data. The proposed algorithm is an extension of the Bayesian quantile chi-squared test proposed by Johnson (2004) for complete data.

Keywords and phrases: Chi-squared test, censoring, goodness-of-fit, survival

8.1 Introduction

Though the use of parametric models, both in a frequentist or Bayesian framework can be advisable in some situations, most applied methods in survival analysis are either non or semi parametric. One of the reasons of the widely use of these methods as opposed to classical parametric approaches relies in the difficulty of checking the validity of a parametric model when data are censored. In this work we propose a Bayesian chi-squared test of goodness-of-fit for censored data. The proposed algorithm is an extension of the Bayesian quantile chi-squared test proposed by Johnson (2004) for complete data.

Pearson's chi-squared test is one of the most classical test of goodness-of-fit. As it is well known the application of this test stands on a finite partition of the sample space in r classes and in the discrepancy between the observed and the expected frequencies in each member of the partition under the null hypothesis. The distribution of Pearson's chi-squared statistic under simple null hypothesis is a chi-square with $r-1$ degrees of freedom. However, when the parameter is unknown (composite hypothesis) and must be estimated, the asymptotic distribution of Pearson's statistic depends on the estimation method. In particular, if the parameter is estimated by maximum likelihood from the complete data

the limit distribution of the statistic is no longer chi-squared. To avoid this difficulty some modifications of the classical Pearson's test have been proposed (see Greenwood and Nikulin, 1996, for a detailed review of chi-squared testing methods).

In the following two sections we describe a Bayesian goodness-of-fit test based on a quantile chi-squared statistic and propose an iterative algorithm to perform this test when dealing with interval–censored survival data.

8.2 Bayesian chi-squared statistic

Let $X_1, \ldots, X_n$ be a random sample from a random variable X. We wish to test the hypothesis H_0 that the distribution of X is $F_0(x;\theta)$ with unknown parameter $\theta = (\theta_1, \ldots, \theta_s)' \in \Theta \subset \mathbf{R}$.

A modification of Pearson's goodness-of-fit test are the chi-squared tests based on sample quantiles (Greenwood and Nikulin, 1996). There are different versions of quantile tests but we describe here the one proposed in Johnson's paper:

Instead of considering a finite partition of the sample space, the quantile test fixes a vector $(p_1, \ldots, p_r)$ of probabilities such that $\sum_{j=1}^{r} p_j = 1$ and $r > s+1$. Define $u_j = p_1 + \cdots + p_i$ for $j = 1, \ldots, r-1 > s$. For each value u_i compute the inverse distribution function $F_0^{-1}(u_j;\theta)$ which define a partition of the sample space: $A_1 = (-\infty, F_0^{-1}(u_1;\theta)]$, $A_2 = (F_0^{-1}(u_1;\theta), F_0^{-1}(u_2;\theta)], \ldots, A_r = (F_0^{-1}(u_{r-1};\theta), +\infty)$. The quantile test statistic is given by

$$X_n^2(\theta) = \sum_{j=1}^{r} \frac{(m_j(\theta) - np_j)^2}{np_j} \tag{8.1}$$

where $m_j(\theta)$ is the number of observations that fall into the jth class, A_j.

An alternative to estimate the unknown parameter by maximum likelihood is to use its posterior distribution in a Bayesian parametric framework. Specifically, Johnson (2004) proposed the following Bayesian quantile statistic:

$$X_n^2(\tilde{\theta}) = \sum_{j=1}^{r} \frac{(m_j(\tilde{\theta}) - np_j)^2}{np_j} \tag{8.2}$$

which corresponds to statistic $X_n^2(\theta)$ defined in (8.1) evaluated at a value $\tilde{\theta}$ of the posterior distribution.

Using results on large sample properties of posterior distributions given in Chen (1985), Johnson (2004) proved that, under the null hypothesis, the asymptotic distribution of the Bayesian quantile statistic $X_n^2(\tilde{\theta})$ is a chi-square distribution with $r-1$ degrees of freedom, independently of the dimension of the underlying parameter vector.

Using this result, Johnson proposed to perform the following goodness-of-fit approach: Obtain a sample $\tilde{\theta}_1, \ldots, \tilde{\theta}_M$ from the posterior distribution of θ. For each of these values compute the proposed statistic and check if the sample $X_n^2(\tilde{\theta}_1), \ldots, X_n^2(\tilde{\theta}_M)$ is consistent with a chi-square distribution with $r-1$ degrees of freedom.

8.3 Bayesian chi-squared test for censored data

In a survival study the random variable of interest X is positive and represents the time until the occurrence of a certain event. We consider the more general situation of interval censoring which contains right and left censoring as special cases. In this framework the potential survival times of n items or individuals, namely, $X_1, \ldots, X_n$, cannot be observed and, instead, we observe intervals that contain them. Let $\mathcal{D} = \{[L_i, R_i], \quad 1 \leq i \leq n\}$ be the interval–censored survival data where L_i is the last observed time for the i^{th} individual before the event has occurred and R_i indicates the first time the event has been observed.

If censoring occurs noninformatively (Gómez *et alt.*, 2004 and Oller *et alt.*, 2004) inferences can be based on the likelihood function $L(\theta|\mathcal{D})$ given by

$$L(\theta|\mathcal{D}) = \prod_{i=1}^{n} \int_{L_i}^{R_i} f_0(u;\theta)\, du. \tag{8.3}$$

To test the null hypothesis, $H_0 : F_X = F_0(\cdot\,;\theta)$, we propose a three steps iterative algorithm. The first two steps correspond to the data-augmentation algorithm proposed in Calle (2003) to obtain a sample from the posterior distribution of the parameter of interest: In the first step, a survival time is sampled for each individual with the restriction that the event occurred between L_i and R_i. We denote by T_i the imputed survival time to distinguish it from the real unobserved survival time X_i of individual i. In the second step the parameter θ is updated based on the complete imputed sample. In the third step the proposed Bayesian quantile statistic (8.2) is computed based on the imputed sample. We denote this statistic by $Y_n^2(\tilde{\theta})$ in order to distinguish it from $X_n^2(\tilde{\theta})$ which would be computed from the real survival times, $X_1, \ldots, X_n$.

The proposed iterative algorithm is as follows:

1. For every $i = 1, \ldots, n$, impute a value T_i sampled from $F_0(x;\theta)$ truncated in the interval $[L_i, R_i]$, that is,

$$f_{T|L,R}(t|l,r) = \frac{f_0(t;\theta)}{F_0(r;\theta) - F_0(l;\theta)} \mathbf{1}_{\{t:\ t\in[l,r]\}}(t)$$

   We obtain an imputed sample $T_1, \ldots, T_n$.

2. Sample a new value $\tilde{\theta}$ of θ from its full conditional distribution given the complete imputed sample $T_1, \ldots, T_n$:

$$p(\theta|T_1, \ldots, T_n) = \prod_{i=1}^{n} f_0(T_i; \theta) \cdot \pi(\theta) \tag{8.4}$$

   where $\pi(\theta)$ is the prior distribution for θ

3. Given the imputed sample, compute the statistic

$$Y_n^2(\tilde{\theta}) = \sum_{j=1}^{r} \frac{(\tilde{m}_j(\tilde{\theta}) - np_j)^2}{np_j} \tag{8.5}$$

   where $\tilde{\theta}$ is the sampled value of θ obtained in step 2 and $\tilde{m}_j$ is the number of imputed values $T_1, \ldots, T_n$ that fall into the $j$the class.

After performing iteratively the above algorithm and after a burn-in process of discarding the first sampled values one obtains a sample $Y_{1n}^2(\tilde{\theta}), \ldots, Y_{Kn}^2(\tilde{\theta})$ of statistic $Y_n^2(\tilde{\theta})$ that is the base for testing the null hypothesis as is described at the end of this section.

The following propositions justify the use of statistic $Y_n^2(\tilde{\theta})$ as a goodness-of-fit test for the distribution of X and give its asymptotic distribution:

Proposition 8.3.1

1. *Under the null hypothesis, $H_0 : F_X = F_0(\cdot\,; \theta)$, the marginal distribution of the imputed values, T_i, is $F_T(t) = F_0(t\,; \theta)$*

2. *Under an alternative hypothesis, $H_1 : F_X = F_1(\cdot\,; \gamma)$, the marginal distribution of the imputed values, T_i, is*

$$F_T(t) = F_0(t; \theta) \iint_{\{(l,r):\ t \in [l,r]\}} \frac{F_1(r; \gamma) - F_1(l; \gamma)}{F_0(r; \theta) - F_0(l; \theta)}\ f_{L,R|X}(l, r|t)\ dlr$$

Proposition 8.3.2 *(Corollary) Under the null hypothesis, H_0, statistic $Y_n^2(\tilde{\theta})$ follows a chi-square distribution with $r - 1$ degrees of freedom as $n \to \infty$.*

As mentioned before, the goodness-of-fit test is based on the sample $Y_{1n}^2(\tilde{\theta}), \ldots, Y_{Kn}^2(\tilde{\theta})$ of the quantile chi-squared statistic $Y_n^2(\tilde{\theta})$, assessing if the sampling distribution agrees with a chi-square distribution with $r - 1$ degrees of freedom. This agreement could be checked in different ways. Johnson (2004) proposed to base the decision on a comparison between the posterior mean and the asymptotic theoretical mean, which in this case is $r - 1$, the degrees of freedom of the asymptotic chi-square distribution.

The performance of the proposed approach has been investigated through a simulation study where the null hypothesis of an exponential distribution was tested for different underlying distributions and different censoring levels.

References

1. Calle, M.L. (2003). Parametric Bayesian Analysis of Interval–Censored and Doubly–Censored Survival Data. *Journal of Probability and Statistical Science* **1**, 103–118

2. Chen, C.F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society (B)*, **47**, 540–546

3. Gómez, G., Calle, M.L. and Oller, R. (2004). Frequentist and Bayesian approaches for interval–censored data. *Statistical Papers*, **45**, pp 139–173

4. Greenwood, P.E. and Nikulin, M.S. (1996). *A Guide to Chi-Squared Testing*, Wiley

5. Johnson, V.E. (2004). A Bayesian chi-squared test for goodness-of-fit, Department of Biostatistics, University of Michigan, http://www.bepress.com/umichbiostat/paper1

6. Oller, R. and Gómez, G., Calle, M.L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood, *The Canadian Journal of Statistics*, **32**, 315–325

9

Discrepancy-Based Model Selection Criteria Using Cross Validation

Joseph E. Cavanaugh, Simon L. Davies, and Andrew A. Neath

Department of Biostatistics, The University of Iowa
Pfizer Development Operations, Pfizer, Inc.
Department of Mathematics and Statistics, Southern Illinois University

Abstract:

A model selection criterion is often formulated by constructing an approximately unbiased estimator of an expected discrepancy, a measure that gauges the separation between the true model and a fitted approximating model. The expected discrepancy reflects how well, on average, the fitted approximating model predicts "new" data generated under the true model. A related measure, the estimated discrepancy, reflects how well the fitted approximating model predicts the data at hand.

Generally, a model selection criterion consists of a goodness-of-fit term and a penalty term. The natural estimator of the expected discrepancy, the estimated discrepancy, corresponds to the goodness-of-fit term of the selection criterion. However, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. It therefore serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the penalty term of the selection criterion.

Cross validation provides a technique for developing an estimator of an expected discrepancy which need not be bias adjusted. The basic idea is to construct an empirical discrepancy that measures the adequacy of an approximating model by assessing how accurately each case-deleted fitted model predicts the deleted case.

The preceding approach is conveniently illustrated in the linear regression framework by formulating estimators of the expected discrepancy based on Kullback's I-divergence and the Gauss discrepancy. The traditional criteria that arise by augmenting the estimated discrepancy with a bias adjustment are the Akaike information criterion and Mallows' conceptual predictive criterion. The corresponding cross-validitory criteria compare favorably to their traditional counterparts in simulation studies.

Keywords and phrases: Akaike information criterion, Mallows' Cp, PRESS

9.1 Introduction

A model selection criterion is often formulated by constructing an approximately unbiased estimator of an expected discrepancy, a measure that gauges the separation between the true model and a fitted approximating model. The natural estimator of the expected discrepancy, the estimated discrepancy, corresponds to the goodness-of-fit term of the selection criterion.

The expected discrepancy reflects how well, on average, the fitted approximating model predicts "new" data generated under the true model. On the other hand, the estimated discrepancy reflects how well the fitted approximating model predicts the data at hand. By evaluating the adequacy of the fitted model based on its ability to recover the data used in its own construction, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. Thus, the estimated discrepancy serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the penalty term of the selection criterion.

Cross validation provides a technique for developing an estimator of an expected discrepancy which need not be bias adjusted. The basic idea involves constructing an empirical discrepancy that measures the adequacy of an approximating model by assessing how accurately each case-deleted fitted model predicts the deleted case.

Cross validation facilitates the development of model selection procedures based on predictive principles. In this work, we attempt to better establish the connection between cross validation and traditional discrepancy-based model selection criteria, such as the Akaike information criterion and Mallows' conceptual predictive statistic.

9.2 Framework for Discrepancy-Based Selection Criteria

Suppose we have an n-dimensional data vector

$$y = (y_1, ..., y_n)',$$

where the y_i's may be scalars or vectors and are assumed to be independent. A parametric model is postulated for y.

Let $F(y)$ denote the joint distribution function for y under the generating or "true" model, and let $F_i(y_i)$ denote the marginal distribution for y_i under

this model. Let $G(y, \theta)$ denote the joint distribution function for y under the candidate or approximating model.

A *discrepancy* is measure of disparity between $F(y)$ and $G(y, \theta)$, say $\Delta(F, G)$, which satisfies

$$\Delta(F, G) \geq \Delta(F, F).$$

We will consider discrepancies of the following form:

$$\Delta(F, G) = \Delta(\theta) = \sum_{i=1}^{n} \mathrm{E}_{F_i} \left\{ \delta_i(y_i; \theta) \right\}.$$

Let $\widehat{\theta}$ denote an estimator of θ. The *overall discrepancy* results from evaluating the discrepancy between $F(y)$ and $G(y, \theta)$ at $\theta = \widehat{\theta}$:

$$\Delta(\widehat{\theta}) = \sum_{i=1}^{n} \mathrm{E}_{F_i} \left\{ \delta_i(y_i, \theta) \right\} |_{\theta = \widehat{\theta}}.$$

The *expected (overall) discrepancy* results from averaging the overall discrepancy over the sampling distribution of $\widehat{\theta}$:

$$\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) \right\} = \sum_{i=1}^{n} \mathrm{E}_F \left\{ \mathrm{E}_{F_i} \left\{ \delta_i(y_i, \theta) \right\} |_{\theta = \widehat{\theta}} \right\}.$$

The *estimated discrepancy* is given by

$$\widehat{\Delta}(\widehat{\theta}) = \sum_{i=1}^{n} \delta_i(y_i, \widehat{\theta}).$$

Model selection criteria are often constructed by obtaining a statistic that has an expectation which is $\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) \right\}$ (at least approximately).

9.3 The Bias Adjustment Approach to Developing a Criterion

The overall discrepancy $\Delta(\widehat{\theta})$ is not a statistic since its evaluation requires knowledge of the true distribution $F(y)$. The estimated discrepancy $\widehat{\Delta}(\widehat{\theta})$ is a statistic and can be used to estimate the expected discrepancy $\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) \right\}$. However, $\widehat{\Delta}(\widehat{\theta})$ is a biased estimator.

Consider writing $\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) \right\}$ as follows:

$$\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) \right\} = \mathrm{E}_F \left\{ \widehat{\Delta}(\widehat{\theta}) \right\} + \left[\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) - \widehat{\Delta}(\widehat{\theta}) \right\} \right].$$

The bracketed quantity on the right is often referred to as the *expected optimism* in judging the fit of a model using the same data as that which was used to construct the fit. The expected optimism is positive, implying that $\widehat{\Delta}(\widehat{\theta})$ is a negatively biased estimator of $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$. In order to correct for the negative bias, we must evaluate or approximate the bias adjustment represented by the expected optimism.

There are numerous approaches for contending with the bias adjustment. These approaches include deriving an asymptotic approximation for the adjustment, deriving an exact expression, or obtaining an approximation using Monte Carlo simulation. However, these methods have limitations since their justifications usually require stringent conditions that may restrict the applicability of the resulting criteria: for example, the assumption that the approximating model of interest is correctly specified or overspecified, the assumption that the largest approximating model in the candidate collection is correctly specified or overspecified, the assumption that the true model errors are normally distributed, the assumption that the sample size is large relative to the dimension of parameter vector for the approximating model, etc.

9.4 Cross Validation Approach to Developing a Criterion

We will now introduce a general cross-validitory estimate of the expected discrepancy that need not be bias adjusted.

Let $y[i]$ denote the data set y with the i^{th} case y_i excluded. Let $\widehat{\theta}[i]$ denote an estimator of θ based on $y[i]$.

Recall that the overall discrepancy is defined as

$$\Delta(\widehat{\theta}) = \sum_{i=1}^{n} \mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}}. \tag{9.1}$$

Now consider the following variant of the overall discrepancy:

$$\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n]) = \sum_{i=1}^{n} \mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}[i]}. \tag{9.2}$$

The expected (overall) discrepancy corresponding to (9.1) is given by

$$\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\} = \sum_{i=1}^{n} \mathrm{E}_F\left\{\mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}}\right\};$$

the expected (overall) discrepancy corresponding to (9.2) is given by

$$\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\} = \sum_{i=1}^{n} \mathrm{E}_F\left\{\mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}[i]}\right\}.$$

Under general conditions, it can be established that

$$\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\} \tag{9.3}$$

and

$$\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\} \tag{9.4}$$

are approximately the same (provided that the sample size is not excessively small). Hence, an unbiased estimator of (9.4) is approximately unbiased for (9.3).

The estimated discrepancy

$$\widehat{\Delta}(\widehat{\theta}) = \sum_{i=1}^{n} \delta_i(y_i, \widehat{\theta})$$

is *negatively biased* for (9.3). However, the empirical discrepancy defined as

$$\widehat{\Delta}^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n]) = \sum_{i=1}^{n} \delta_i(y_i, \widehat{\theta}[i]) \tag{9.5}$$

is *exactly unbiased* for (9.4). The justification of this fact is straightforward.

Since $\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\} \approx \mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$, it follows that $\widehat{\Delta}^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])$ is *approximately unbiased* for $\mathrm{E}_F\left\{\widehat{\Delta}(\widehat{\theta})\right\}$. Thus, the empirical discrepancy $\widehat{\Delta}^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])$

(a) estimates $\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\}$ without bias,

(b) estimates $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$ with negligible bias for large n.

The preceding are general results that may be established without imposing restrictive conditions.

9.5 Examples in the Linear Regression Setting

Consider a setting where a continuous response variable is to be modeled using a linear regression model.

Under the approximating model, assume the y_i are independent with mean $x_i'\beta$ and variance σ^2. Let $\theta = (\beta'\sigma^2)'$. Further, let $g(y,\theta)$ denote the approximating density for y, and let $g_i(y_i,\theta)$ denote the approximating density for y_i.

Kullback's I-Divergence and the Gauss discrepancy have applicability to many modeling frameworks, including linear regression. The I-divergence is given by

$$\Delta_{\mathrm{I}}(\theta) = \mathrm{E}_F \left\{-2 \ln g(y, \theta)\right\} = \sum_{i=1}^{n} \mathrm{E}_{F_i} \left\{\delta_i^{\mathrm{I}}(y_i; \theta)\right\}, \tag{9.6}$$

where $\delta_i^{\mathrm{I}}(y_i; \theta) = -2 \ln g_i(y_i, \theta)$. The Gauss (sum of squares) discrepancy is given by

$$\Delta_{\mathrm{G}}(\theta) = \mathrm{E}_F \left\{ \sum_{i=1}^{n} (y_i - x_i^{'} \beta)^2 \right\} = \sum_{i=1}^{n} \mathrm{E}_{F_i} \left\{\delta_i^{\mathrm{G}}(y_i; \theta)\right\}, \tag{9.7}$$

where $\delta_i^{\mathrm{G}}(y_i; \theta) = (y_i - x_i^{'} \beta)^2$.

Provided that the approximating model of interest is correctly specified or overspecified, the Akaike information criterion provides an asymptotically unbiased estimator of the expected discrepancy corresponding to (9.6). Provided that the largest approximating model in the candidate collection is correctly specified or overspecified, a simple variant of Mallows' conceptual predictive statistic (with identical selection properties) provides an exactly unbiased estimator of the expected discrepancy corresponding to (9.7). This variant is given by $(\mathrm{C}_p + n)\mathrm{MSE}_L$, where C_p denotes Mallows' statistic and MSE_L denotes the error mean square for the largest approximating model.

Assuming normal errors, the cross-validitory criterion (9.5) based on the I-divergence (9.6) is given by

$$\sum_{i=1}^{n} \ln \hat{\sigma}_{-i}^2 + \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,-i})^2}{\hat{\sigma}_{-i}^2},$$

where $\hat{y}_{i,-i}$ denotes the fitted value for y_i based on the data set $y[i]$, and $\hat{\sigma}_{-i}^2$ denotes the MLE for the variance based on the data set $y[i]$.

The cross-validitory criterion (9.5) based on the Gauss discrepancy (9.7) is given by

$$\sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2,$$

the well known PRESS (predictive sum of squares) statistic.

In simulation studies, the cross-validitory criteria compare favorably to their traditional counterparts. In settings where the generating model is among the collection of candidate models under consideration, the cross-validitory criteria tend to select the correctly specified model more frequently and to select overspecified models less frequently than their bias-adjusted analogues.

10

A Competing Risks Model for Degradation and Traumatic Failure Times

Vincent Couallier

Equipe Statistique Mathématique et ses Applications
Université Victor Segalen Bordeaux 2 FRANCE
couallier@sm.u-bordeaux2.fr

Abstract: We are interested here in some failure times due to wear or aging. The main aim is to jointly model the degradation process and one (or more) associated failure time(s). Two main joint models exist. The first one considers a failure time which is directly defined by the degradation process (degradation failure) as a hitting time of growth curve with random coefficients, the second one considers that the degradation influences the hazard rate of a failure time by a conditional definition of its survival function (traumatic failure). When both modes of failure exist, only the first one is observed. Very often, longitudinal observations of degradation values (measured with error) are available for each item until the first failure. We are mainly interested here in the nonparametric estimation of the cumulative intensity function of the traumatic failure time and related reliability characteristics. In order to analyze the distribution of the degradation failure, we use either pseudo degradation failures or parametric nonlinear mixed regression model.

Keywords and phrases: Degradation failure time, Traumatic failure time, nonlinear mixed regression, Nelson-Aalen estimator.

10.1 Introduction

Degradation data modeling presents an attractive alternative in the assessment and improvement of reliability of components from which the overall system reliability can be deduced. If a component is monitored during its operation time, periodical tests can provide either the simple information that the component performs well (and thus is at risk for a failure) or a quantitative information giving the level of degradation in a specified scale at every time measurement. Thus the degradation process can sometimes be observed and monitored through some quantitative characteristics. Examples of such degra-

dation characteristics for monitoring degradation processes include the wear of tires (de Oliveira and Colosimo, 2004), gain of transitors (Whitmore, 1995), degradation of fluorescent lamps (Tseng, Hamada and Chiao, 1995) or catalytic converters for automotive (Barone, M. Guida, G. Pulcini, 2001) among others.

The usual traumatic failure time has then to be related to the evolution of the degradation process. Two main joint models exist. The first one considers a failure time which is directly defined by the degradation process, the second one considers that the degradation process influences the distribution of the failure time through a conditional definition of its hazard rate.

Let us assume that the degradation of an item is given by the sample path of a non decreasing real-valued right continuous and left hand limited stochastic process $Z(t)$, $t \in I$. Lawless and Crowder (2004) and Couallier (2004) consider gamma processes, Kahle and Wendt (2004) and Lehmann (2004) consider marked point processes and Whitmore and Schenkelberg (1997) consider Wiener diffusion processes. In the following, we shall make the assumption that

$$Z(t) = \mathcal{D}(t, A), t > 0, \tag{10.1}$$

where $\mathcal{D}$ is a differentiable and non decreasing parametric function of the time and A is a random variable in $\mathbb{R}^p$ which takes account on the variability of the degradation evolution. The model reduces here to a nonlinear growth curve model with random coefficients where, for each individual $i = 1..n$ the unknown real degradation is $Z^i(t) = \mathcal{D}(t, A_i)$ where A_i is the realization of A for the i-th item and the observed degradation values are

$$Z_j^{i|obs} = \mathcal{D}(t_{ij}, A_i) + \epsilon_j^i,$$

measured at times $t_{ij}, j = 1..n_i$ where the ϵ_j^i are error measurements of the degradation values.

10.2 The degradation failure - estimation of F_A and F_{T_0}

We assume that the life time T_0 is the first time of crossing a fixed ultimate threshold z_0 for Z(t)

$$T_0 = \inf\{t \in I, Z(t) \geq z_0\}.$$

The failure time T_0 is sometimes called soft failure (or failure directly due to wear) because in most of industrial applications, z_0 is fixed and the experiment is voluntarily ceased at the time the degradation process reaches the level z_0 or just after this time. Known results about parametric models of degradation failure Time T_0 give the distribution function of T_0 with respect to the distribution function od $Z(t)$ or A. For instance, Padgett and Tomlinson (2004) use the fact that if Z is a gaussian process with positive drift then T_0 follows

an Inverse Gaussian distribution, De Oliveira and Colosimo (2004) assume the path model $Z(t) = a + bt$ where a is fixed (unknown) and b is Weibull(α, β). Then T_0 follows a Inverse Weibull whose parameters depend on z_0, a, α and β. Yu (2003) assume the decreasing path model $Z(t) = -\beta t^{\alpha}$ where α is fixed and $\beta \sim LN(\mu, \sigma^2)$ then $T_0 \sim LN(\ (\ln(-z_0) - \mu)/\alpha, \sigma^2/\alpha^2)$.

As an example, we shall analyze in the following twenty one degradation curves describing the fatigue crack propagation in aluminium alloy materials (Meeker and Escobar, 1998). Each curve is well fitted by a Paris Curve with a high variability in the adjusted parameters. The Paris growth curve is given here by

$$g(t, m, C) = \left(0.9^{\frac{2-m}{2}} + \frac{2-m}{2} C \sqrt{\pi}^{m} t\right)^{\frac{2}{2-m}},$$

with unit-to-unit coefficients $A^i = (m_i, C_i)$ fitted on each item. The failure due to degradation is defined as the time where the curve reach the threshold $z_0 = 1.6$. The aim is thus to estimate the distribution functions $F_{(m,C)}$ and F_{T_0} with the noised measurements of degradation for each item without assuming that $F_{(m,C)}$ lies in a parametric family of distribution functions. For purely parametric estimation of degradation curves with maximum likelihood estimation of the d.f. of the failure time T_0 only due to wear, we refer to Meeker and Escobar (1998) and reference therein. Comparison with our semiparametric model will be provided.

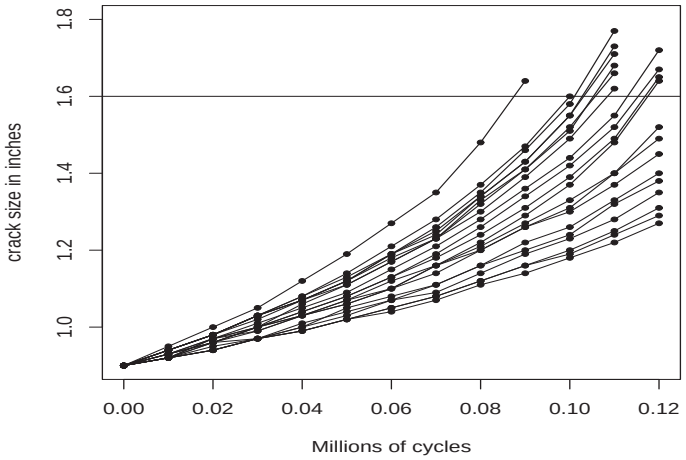


Figure 10.1: Fatigue crack size propagation for Alloy-A data

Each individual path leads to a prediction $(\hat{A}_i)$ of unknown (A_i) by non linear least squares method. For all i, a predictor $\hat{A}^i$ is computed with nonlinear least squares method with observed degradation values $Z_{ij}^{obs}, j = 1..n_i$. Bagdonavicius and Nikulin (2004) have shown under technical assumptions that the pseudo empirical cumulative distribution function $\hat{\hat{F}}_A(a) = 1/n \sum_{i=1}^{n} \mathbf{1}_{\{\hat{A}^i \leq a\}}$ is

a uniformly consistent estimator of F_A as $n \to +\infty$. Instead of plugging the $\hat{A}^i$ in the unknown empirical measure $P(E) = 1/n \sum_{i=1}^{n} 1(A^i \in E)$, we propose here to use the approximate distribution function of $\hat{A}^i$ around A^i which is gaussian with mean zero and estimated variance matrix $\hat{\Sigma}_i$ given by the numerical least square method. If, for all i, $\hat{A}^i - A^i \sim \mathcal{N}(0, \hat{\Sigma}_i)$ then a estimator of the cumulative distribution function F_A is

$$\tilde{F}_A(a) = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^p} 1_{(u<a)} f_{\mathcal{N}(\hat{A}^i, \hat{\Sigma}_i)}(u) du \tag{10.2}$$

Marginal distributions are easily deduced. For each coordinate A_k of A, the estimated cumulative distribution function is

$$\tilde{F}_{A_k}(a) = \frac{1}{n} \sum_{i=1}^{n} \Phi(\frac{a - \hat{A}_k^i}{\hat{\sigma}_k^i})$$

where $\hat{\sigma}_k^{2i}$ is the estimated variance of $\hat{A}_k^i$ and Φ is the cumulative distribution function of the standard normal law.

The distribution function of T_0 is obtained either by calculating the pseudo-failure times $\hat{T}_0^i = h(z_0, \hat{A}_i)$ and plugging it in the unknown empirical cumulative distribution function of $T_{0i},\ i = 1..n$ or by using $P(T_0 < t) = P(\mathcal{D}(t, A) \geq z_0)$ and

$$\hat{F}_{T_0}(t) = \int 1_{\{\mathcal{D}(t,a) \geq z_0\}} d\tilde{F}_A(a)$$

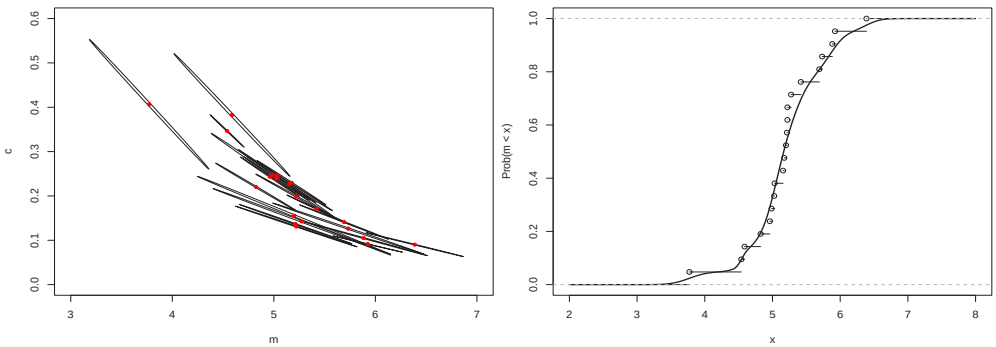


Figure 10.2: Predictors $(\hat{m}_i, \hat{c}_i)$ of (m_i, c_i) for Alloy-A Data and 95% confidence region for each estimation – ecdf of predicted $\hat{m}_i$ and $\tilde{F}_m$

10.3 A joint model with both degradation and traumatic failure times

As in Bagdonavicius and Nikulin (2004) and Couallier (2004), we define the traumatic failure time T with the conditional survival function given the past degradation process as

$$P(T > t|Z(s), 0 \leq s \leq t) = \exp\left(-\int_0^t \lambda_T(Z(s))ds\right). \tag{10.3}$$

λ_T is a non decreasing function living in the degradation domain. The aim is to estimate this failure rate which depends on the chronological time only through the degradation process. The higher the degradation is, the higher the instantaneous probability of failure for an at-risk item will be. Also, the conditional survival function depends on the whole past degradation process. In this model, contrarily to T_0, the traumatic failure time T can occur even if the degradation level is low but its survival function depends on the degradation function.

We assume that T and T_0 are two competing failure times whose distribution functions are related to the degradation process. $U = \min\{T, T_0\}$ is the observed failure time. If $U = T_0$, we do not observe the traumatic failure time T. The function $\Lambda(z) = \int_0^z \lambda(s)ds$ is the cumulative hazard in the degradation space. The definition (10.3) reduces here to

$$R_T(t|A = a) = P(T > t|A = a) = \exp\left(-\int_0^t \lambda(\mathcal{D}(s, a))ds\right)$$

For each item $i = 1..n$, by denoting $T_0^i = \inf\{j \in \{1..n_i\}|Z_j^i \geq z_0\}$, we observe only $U^i = \min(T^i, T_0^i, t_{n_i}^i)$ and $\delta^i = 1(U^i = T^i)$ where $t_{n_i}^i$ is the last time of observation. In order to get nonparametric estimates of Λ, of $R_T(t|A = a)$ and $R_T(t) = E_A(R(t|A))$, we use the fact that, denoting $h(., A)$ the inverse function of $\mathcal{D}(., A)$

$$R_T(t|Z(s), 0 \leq s \leq t) = P(T > t|A) = \exp\Big[-\int_{\mathcal{D}(0,A)}^{\mathcal{D}(t,A)} h'(z, A)d\Lambda(z)\Big]$$

If we denote by Z_i the last observed degradation value (reached at time U_i), a Doob Meyer decomposition of some counting process in the degradation space leads to a nonparametric estimator of Λ

$$\hat{\Lambda}(z) \quad = \quad \sum_{\delta_i=1, Z_i \leq z} \Big(\frac{1}{\sum_{j, Z_j \geq Z_i} h'(Z_i, \hat{A}^j)}\Big),$$

and a nonparametric estimator of the cond. survival function is

$$\hat{R}_T(t|A) = \exp\Big[-\int_{\mathcal{D}(0,A)}^{\mathcal{D}(t,A)} h'(z, A)d\hat{\Lambda}(z)\Big].$$

Estimation of the overall survival function needs integration w.r.t. F_A. A nonparametric estimator of $R_T(t) = P(T > t)$ is

$$\hat{R}_T(t) \quad = \quad \int \exp\Big[-\int_{\mathcal{D}(0,A)}^{\mathcal{D}(t,A)} h'(z, A)d\hat{\Lambda}(z)\Big]d\tilde{F}_A.$$

Estimation of the survival function of $U = min(T, T_0)$ is also available.

References

1. Bagdonavicius, V., Nikulin, M. (2000). Estimation in degradation Models with Explanatory Variables, Lifetime Data Analysis, **7** : 85-103.
2. Bagdonavicius, V., Nikulin, M. (2004). Semiparametric analysis of degradation and failure time data with covariates, in *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, Birkauser.
3. Barone,S., Guida,M., Pulcini,G. (2001). A stochastic degradation model of catalytic converters performances, in *MECA'01 Intern. Workshop on Modeling and Control in Automotive Engines*, technical paper.
4. Couallier, V. (2004). Comparison of parametric and semiparametric estimates in a degradation model with covariates and traumatic censoring in *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, Birkauser.
5. Kahle, W., Wendt H. (2004). On a cumulative damage process and resulting first passages times. *Applied Stochastic Models in Business and Industry*, **20**(1), 17-26
6. Lawless J, Crowder M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, **10**(3), 213-27.
7. Lehmann, A. (2004). On a Degradation-Failure Model for Repairable Items, In *Semiparametric Models and Applications to Reliability, Survival Analysis and Quality of Life.* (Ed., M. Nikulin, M. Mesbah, N. Balakrishnan and N. Limnios), Birkhäuser, Boston.
8. Meeker, W.Q. and Escobar, L. (1998). *Statistical Analysis for Reliability Data*, John Wiley and Sons, New York.
9. de Oliveira, V.R.B. and Colosimo, E.A. (2004), Comparison of Methods to Estimate the Time-to-failure Distribution in Degradation *Tests. Qual. Reliab. Engng. Int.*, **20** 363-373.
10. Padgett, W.J. and Tomlinson, M.A. (2004). Inference from Accelerated Degradation and Failure Data Based on Gaussian Process Models, *Lifetime Data Analysis*, **10**, 191-206.
11. Tseng S.T., Hamada M., Chiao C.H. (1995). Using degradation data from a factorial experiment do improve fluorescent lamp reliability. *Journal of Quality Technology*, **27** 363-369.
12. Whitmore, G.A.(1995). Estimating degradation by a wiener diffusion process subject to measurement error. *Lifetime Data Anal.*,**1**, 307-319.
13. Whitmore, G.A. and Schenkelberg, F. (1997). Modelling accelerated degradation data using Wiener diffusion with a time scale transformation. *Lifetime Data Analysis*, **3**, 27-45.
14. Yu, H.F. (2003). Designing an accelerated degradation experiment by optimizing the estimation of the percentile. *Quality and Reliability Engineering International*, **19**, 197-214.

11

Measuring Degradation of Pollution Related Quality of Life in the SEQAP Study

Deguen S*, Segala C and Mesbah M*****

**LAPSS, Ecole Nationale de Sante Publique, Rennes*
***SEPIA-Sante, Melrand*
****LSTA, Universite Pierre et Marie Curie, Paris*

This work is funded by ADEME in the framework of Primequal-Predit.

Abstract: In this work, using a real epidemiological survey, we will mainly present the methodology of construction of a Quality of Life questionnaire specific to air pollution disturbance.

Keywords and phrases: Quality of Life, Air Pollution, Generalized Linear Models, Item Response Theory,Unidimensionality, Clustering of Variables

11.1 Introduction

Air pollution may cause cardio-respiratory diseases, and more often annoyance reactions. Despite the large populations exposed to air pollution in our cities and numerous epidemiological study demonstrating relationships between air pollution and health, few studies have been published on the quantitative relations between the exposure to pollution and the public perception of air quality. The SEQAP epidemiological study has for main objective to measure the relationships between adults perception of air pollution and air pollutants concentrations measured by monitoring networks in several French towns. Around 3 000 subjects will be randomly selected from adults living in 7 cities having different levels of air pollutants exposure. From each cities, 450 subjects aged 25-65 will be chosen. Interview will be conducted by phone, including questions on socio-demographic characteristics, occupation, smoking habits, household members, access to a car, health, plus a specific quality of life scale taking into account air pollution annoyance.

In this work we will mainly present the methodology of construction of a Quality of Life scale specific to air pollution disturbance.

11.1.1 Finding Questions

During a preliminary step, the main goal was to answer the question: what do we want to measure? We found only few bibliographical references on the subject. Unlike most of the studies on Perception of Air Pollution mainly based on assessment of Satisfaction about Air Quality, we focused on assessment of degradation of Quality of Life explained by air pollution. The first step was to found questions (qualitative items) related to that subjective concept. These questions (items) were chosen using a preliminary deep bibliographical research and four focus group meetings. Two different focus groups involved students in Environmental health, another one included teachers, known as expert on Health Environment and the last one included general people without any a priori knowledge on Environmental science.

After this preliminary step, we get a form containing questions on annoyance reactions for different fields: health, daily life, local environment and quality of life.

11.1.2 Selecting Questions

The second step consisted on testing this questionnaire on a small group of 83 subjects. All interviews were done by telephone. In order to get a preliminary sample including people living in places with contrasting levels of air pollution three different cities were chosen. 26 interviews were obtained from people living in Le Havre, 16 inhabitants of Lyon and 41 from people living in Rennes. We present in this paper preliminary results of the analysis of the obtained data. The main interest of this preliminary study is to test the acceptability of the questionnaire and to eliminate very bad questions. The final validation study, and the selection of items will be based on the data of the large main survey, available later.

11.2 Classical Unidimensional Psychometric Models

Statistical validation methods are mainly based on psychometric unidimensional models.

11.2.1 The parallel model describing the unidimensionality of a set of variables

Let X_1, X_2, ..., X_k, a set of observed variables measuring the same underlying unidimensional latent (unobserved) variable. We define X_{ij} as the measurement of subject i, i=1,..,n, given by a variable j, where j=1,..,k. The model underlying Cronbach's Alpha is just a mixed one-way anova model: $X_{ij} = \mu_j + \alpha_i +$

ε_{ij}, where μ_j is a varying fixed (non-random) effect and α_i is a random effect with zero mean and standard error σ_α corresponding to subject variability. It produces the variance of the true latent measure ($\tau_{ij} = \mu_j + \alpha_i$). ε_{ij} is a random effect with zero mean and standard error σ corresponding to the additional measurement error. The true measure and the error are uncorrelated: $cov(\alpha_i, \varepsilon_{ij}) = 0$. This model is called parallel model, because the regression lines relating any observed item X_j, j=1,...,k and the true unique latent measure τ_j are parallel.

These assumptions are classical in experimental design. This model defines relationships between different kinds of variables: the observed score X_{ij}, the true score τ_{ij} and the error ε_{ij}. It is interesting to make some remarks about assumptions underlying this model. The random part of the true measure of individual i is the same whatever might be variable j. $\alpha_i$does not depend on j. The model is unidimensional. One can assume that in their structural part all variables measure the same thing (α_i).

11.2.2 Reliability of an instrument

A measurement instrument gives us values that we call observed measure. The reliability ρ of an instrument is defined as the ratio of the true over the observed measure. Under the parallel model, one can show that the reliability of any variable X_j (as an instrument to measure the true value) is given by:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2}$$

which is also the constant correlation between any two variables. This coefficient is also known as the intra-class coefficient. The reliability coefficient ρ can be easily interpreted as a correlation coefficient between the true and the observed measure.

When the parallel model is assumed, the reliability of the sum of k variables equals:

$$\tilde{\rho} = \frac{k\rho}{k\rho + (1 - \rho)}$$

This formula is known as the Spearman-Brown formula. Its maximum likelihood estimator, under the assumption of a normal distribution of the error and the parallel model, is known as Cronbach's Alpha Coefficient (CAC) [Cronbach (1951)]:

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum_{j=1}^{k} S_j^2}{S_{tot}^2})$$

where $S_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \overline{X_j})^2$ and $S_{tot}^2 = \frac{1}{nk-1} \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \overline{X})^2$.

11.2.3 Backward Cronbach Alpha Curve

The Spearman-Brown formula indicates a simple relationship between CAC and the number of variables. It is easy to show that the CAC is an increasing function of the number of variables. This formula is obtained under the parallel model.

A step-by-step curve of CAC can be built to assess the unidimensionality of a set of variables. The first step uses all variables to compute CAC. Then, at every successive step, one variable is removed from the scale. The removed variable is that one which leaves the scale with its maximum CAC value. This procedure is repeated until only two variables remains. If the parallel model is true, increasing the number of variables increases the reliability of the total score which is estimated by Cronbach's alpha. Thus, a decrease of such a curve after adding a variable would cause us to suspect strongly that the added variable did not constitute a unidimensional set with the other variables.

11.3 Modern measurement models and graphical modeling

Modern ideas about measurement models are more general. Instead of arbitrarily defining the relationship between observed and truth as an additive function (of the true and the error), they just focus on the joint distribution of the observed and the true variables f(X,gθ). We do not need to specify any kind of distance between X and θ. E and its relation to X and θ could be anything! E is not equal to X-gθ. E could be some kind of distance between the distributions of X and θ.

This leads us naturally to Graphical Modelling, as presented briefly in the introduction of this paper. Graphical modelling (Lauritzen and Wermuth, (1989), Whittaker, (1990)) aims to represent the multidimensional joint distribution of a set of variables by a graph. We will focus on conditional independence graphs. The interpretation of an independence graph is easy. Each multivariate distribution is represented by a graphic, which is composed of nodes and edges between nodes. Nodes represent one-dimensional random variables (observed or latent, i.e., non-observed) while a missing edge between two variables means that those two variables are independent conditionally on the rest (all other variables in the multidimensional distribution).

The Rasch Model in the psychometric context is probably the most popular of modern measurement models. It is defined for the outcome X taking two values (coded for instance 0 or 1):

$$P(X_{ij} = 1/\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1+\exp(\theta_i - \beta_j)}$$

θ_n is the person parameter: it measures the ability of an individual n, on the latent trait. It is the true latent variable in a continuous scale. It is the true score that we want to obtain, after the reduction of the k items to 1. β_j is the item parameter. It characterizes the level of difficulty of the item (the question). The Rasch model is member of the Item Response Models (Fischer and Molenaar (1995)). The Partial Credit Model (Fischer and Molenaar (1995)) is another member of the family of Item Response Model: it is the equivalent to the Rasch Model for ordinal categorical responses. Let $P_{ijx} = P(X_{ij} = x)$, then

$$P_{ijx} = \frac{\exp\left(x\theta_i - \sum_{l=1}^{x} \beta_{jl}\right)}{\sum_{h=0}^{m_j} \exp\left(h\theta_i - \sum_{l=1}^{h} \beta_{jl}\right)},$$

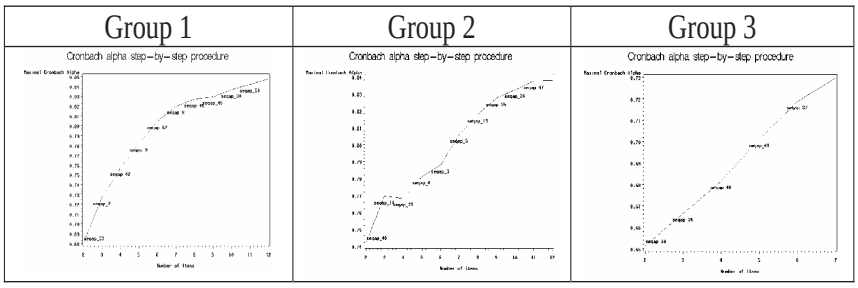
for x= 1. 2, . . ., m_j (m_j is the umber of levels of item j); i = 1 ... N (number of subjects); j= 1 ... k (number of items). Under these models a reliability coefficient like to Cronbach alpha can be derived (Hamon and Mesbah (2002)) and used in the same way as in parallel models, and a Backward Cronbach alpha curve can be used at a first step followed by a goodness of fit test of the Rasch model.

11.4 Results

Fifty four (54) ordinal items were used in the original form to measure the annoyance of air pollution. Four response levels were used: "pas du tout (never); parfois (sometimes);souvent (often);tout le temps (always)" At a first step, nine (9) items with ceiling effects (more than 90 per cent of persons answering "never") were excluded from the analyse. Then a forced (limited to three factors) factorial analysis followed by a varimax rotation allows us to identify three different groups of items. Then a Stepwise Cronbach Alpha Curve was built. Few items were deleted to allow the Cronbach Alpha Curve to be an increasing curve. Final curves are below. A Rasch internal analysis and various external validations using covariate in the questionnaire were also performed. The contents of the items was discussed with psychological and medical expert on respiratory diseases. Giving specific definition to the identified item groups was the first objective of such discussion with psychological and medical experts.

11.5 Discussion

The previous results were based on a small pre-study. These results are only useful to exclude some "very bad" items to get a smaller questionnaire. The



groups found here need to be confirmed by the final validation study based on the planned large study.

References

1. Cronbach LJ. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* vol. 16, pp. 297-334.

2. Fisher,G.H. and Molenaar, I.W. (1995) *Rasch models, Foundations, recent Developments and Applications*, Springer-Verlag, New-York.

3. Hamon A. and Mesbah M. (2002). Questionnaire Reliability under the Rasch Model. In: *Statistical Methods for Quality of Life Studies: Design, Measurement and Analysis.* Eds., M. Mesbah, B.F. Cole, and M.L.T. Lee), pp. 155-168 Kluwer Academic, Boston.

4. Whittaker, J. (1990). *Graphical Models In Applied Multivariate Statistics, first edition* Wiley, New-York, 1990.

5. Lauritzen S.L. and Wermuth N. (1989), Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics.* Vol 17 N ° 1 p31-57.

12

Diagnostic Plots for the Frailty Distribution in Proportional Hazards Models

P. Economou and C. Caroni

Department of Mathematics,
National Technical University of Athens,
9 Iroon Polytechniou, Zografou,
157 80 Athens, Greece
(polikon@math.ntua.gr)

Abstract:

A graphical diagnostic test for the correct choice of frailty distribution is constructed by exploiting a closure property of common frailty distributions in proportional hazards models, namely, that the distribution of frailty among survivors at time t has the same form as the initial distribution, with some parameters unchanged. The distribution of frailty among surviving clusters in the case of shared frailty is obtained under various definitions of the lifetime of a cluster. The test is extended to this situation when cluster lifetime is defined as the shortest lifetime of the cluster's members. Other definitions of cluster lifetime are less useful for this purpose because the distribution of frailty among surviving clusters at time t does not have the same form as the initial distribution.

Keywords and phrases: Lifetime data; frailty; shared frailty; proportional hazards; graphical diagnostics

12.1 Introduction

Heterogeneity between individuals in time-to-event studies may be accounted for by including measured covariates and an unmeasured frailty in the model. In a proportional hazards framework, the model

$$h(t|z;\mathbf{x}) = ze^{\beta'\mathbf{x(t)}}h_b(t) \tag{12.1}$$

is usually assumed, where $\mathbf{x}(.)$ is a vector of possibly time-dependent covariates. The unobservable individual random effect Z is the frailty and h_b is a baseline hazard function.

One extension is to *shared frailty models*, where the structure of the data is such that individuals are in groups or clusters and all the members of the same cluster share the same value of frailty. Thus (omitting covariates, which could be at the individual or cluster level, or both)

$$h_{ij}(t_j|z_i) = z_i h_b(t_j) \tag{12.2}$$

for the $j = 1, ..., m_i$ members of the ith cluster.

A wide range of distributions are available to model the non-negative random variable Z. Common choices include the Gamma (Vaupel et al, 1979) and Inverse Gaussian (Hougaard, 1984), which are both members of a class of exponential family distributions that have an interesting and useful property, as seen below. Whatever distribution is assumed, it is desirable to check that it is supported by the data. The present paper develops graphical diagnostics for this purpose.

12.2 Closure property of the frailty distributions

Let frailty Z be a random variable with distribution $F(z;\alpha)$ on $(0,\infty)$, where α is the parameter vector, with p.d.f. of the form

$$f_Z(z) = \frac{e^{-[z,g(z)][\eta_1(\alpha),\eta_2(\alpha)]'}}{\Phi(\alpha)}\ \xi(z)$$

which is an exponential family distribution with canonical statistics z and $g(z)$ (Shao, 1998). A closure property for this distribution was shown by Hougaard (1984) and earlier for the special case of the gamma distribution by Vaupel (1979). The following theorem extends Hougaard's result by including covariates.

Theorem 12.2.1 *Given the frailty distribution $F(z;\alpha)$ with p.d.f. as above, then under the proportional hazards frailty model the frailty distribution among survivors at time t is again $F(.)$. The value of $\eta_1(\alpha)$, the element of the parameter vector corresponding to z, changes, but the components of $\eta_2(\alpha)$ do not. More specifically, the p.d.f. of frailty among survivors at time t is given by*

$$f_{Z_{|T>t}}(z) = \frac{e^{-[z,g(z)][\eta_1^*(\alpha),\eta_2(\alpha)]'}}{\Phi^*(\alpha)}\ \xi(z)$$

where $\eta_1^*(\alpha) = \eta_1(\alpha) + H_b^{\mathbf{x}}(t)$ and $\Phi^*(\alpha) = \Phi(\alpha)S_T(t)$.

This result is obtained from the joint p.d.f. of T and Z, expressed as the product of the p.d.f. of Z and the conditional p.d.f. of T given $Z = z$.

This class of the exponential family includes the Generalized Inverse Gamma distribution and hence the Gamma and Inverse Gaussian distributions as special cases. Some other useful frailty distributions, such as the lognormal, do not belong to this class because they do not have z as a canonical statistic. This obstacle can be overcome by considering a generalized distribution, adding one more parameter (Hougaard, 1986) which will be zero initially. The practical importance of this property for our purposes is that we can fit the same model (for example, Gamma frailty with Weibull baseline hazard) to a sample of lifetime data, or to a subsample consisting of the survivors at any chosen time, and the estimates of the parameter η_2 should be stable because the corresponding population parameter does not change. In the case of the Gamma distribution, this is the shape parameter. The property has been used by Economou and Caroni (2005) to construct a diagnostic plot for the assumed frailty distribution in the individual frailty model. An example of this is shown below, followed by its extension to the shared frailty model.

12.3 Illustration

The maximum likelihood estimates of the model's parameters are obtained by maximising the logarithm of the usual likelihood function for lifetime data

$$L = \prod_{i=1}^{n} \left\{ h(t_i)^{\delta_i} S(t_i) \right\} \tag{12.3}$$

where δ_i is the censoring indicator which takes the value 1 if t_i is an observed lifetime and zero if it represents a right censored observation. We first carry out this estimation using all the data. Then we select a sequence of convenient time points τ_j $(j = 1, 2...k)$ and repeat the estimation k times, using in the ith estimation only those data points t_i satisfying $t_i \geq \tau_j$.

To illustrate the method, we used data on the duration of a treadmill test undertaken by 978 successive patients at a cardiac clinic in Athens. Figure 12.1 shows the diagnostic plots defined above. The baseline hazard is assumed to be Weibull and both the Gamma and the Inverse Gaussian are examined as possible distributions of the frailty. The upper diagram shows successive estimates of the shape parameter (more precisely, the ratio of successive estimates to the initial estimate) when the Gamma distribution is assumed for frailty. The lower diagram shows the corresponding results when an Inverse Gaussian distribution of frailty is assumed. To assist in assessing the results, an envelope of simulated values has been added (Economou and Caroni, 2005). These diagrams indicate clearly that the assumption of a Gamma distribution for frailty is acceptable, because the estimates of its shape parameter at different times are scattered about a horizontal line, but the Inverse Gaussian assumption is not.

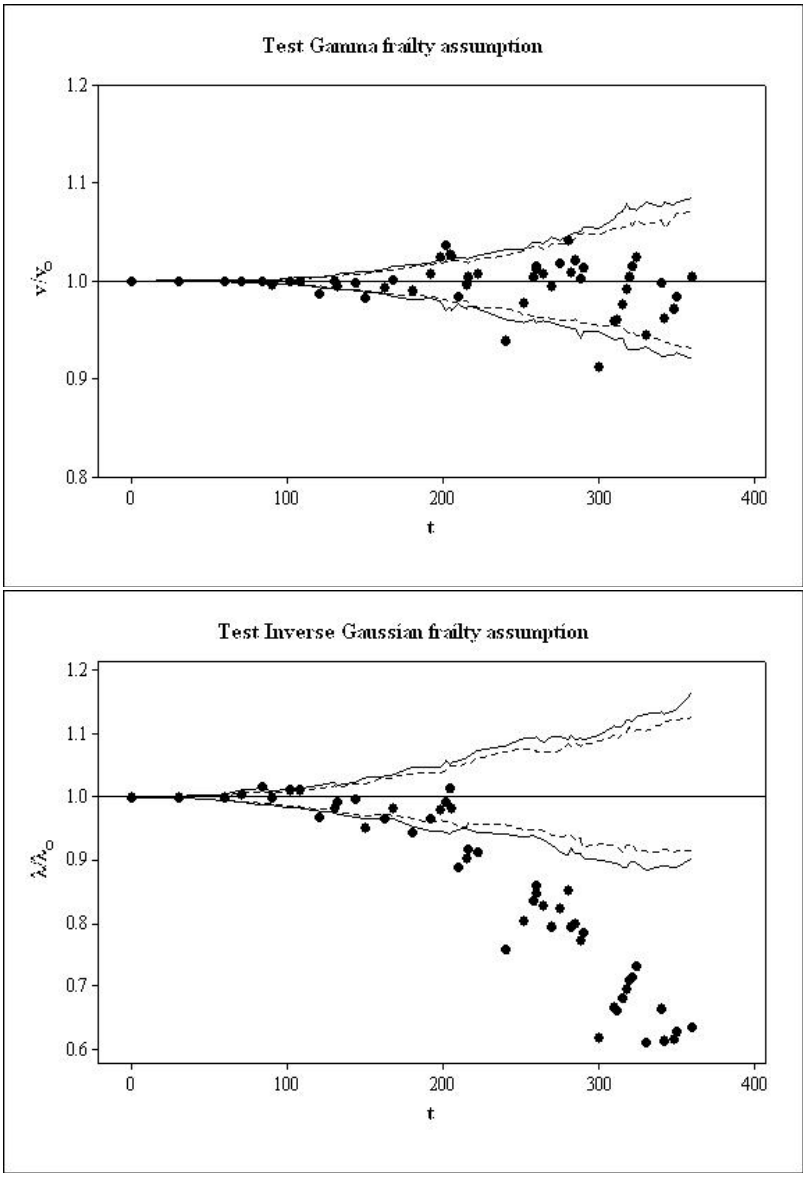


Figure 12.1: Diagnostic plots for data on 978 cardiac patients. Top: Weibull-Gamma mixture (=Burr distribution); bottom: Weibull-Inverse Gaussian mixture.

12.4 Shared frailty

To apply the same idea to the case of shared frailty, it is necessary to define the survivors at time t. Since the distribution of the random variable Z is over the clusters, the unit of analysis will be clusters not individuals. We therefore

need to define what is meant by saying that a cluster survives at time t. Two cases are of immediate interest.

Case 1.
The cluster "dies" as soon as any member dies. Hence the lifetime distribution of a cluster of size m is given by the random variable $T_{(1)} = \min_{i=1,\ldots,m} T_i$. This could be called the "minimum" definition. In this case, since the survivor function of the cluster given frailty z, is $S_{(1)}(t|z) = (S(t|z))^m$, properties at the cluster level are basically the same as those found already for individual frailty. In particular, the closure property applies. It is easy to show that the parameter denoted above as η_1 becomes $\eta_1 + mH_b(t)$ and η_2 remains unchanged. If all clusters have the same size, the previous analysis therefore applies without change. If there are different sizes of clusters, then the overall likelihood is

$$L = \prod_{m=1}^{k} \prod_{i=1}^{n_m} \left\{ h(t_{mi})^{\delta_i} S(t_{mi}) \right\} \tag{12.4}$$

where t_{mi} is the survival time (either observed or censored) of the ith cluster of size m and k is the largest size of cluster. The only difference from the likelihood for a single size is that the η_1 parameter takes the different form indicated above for each size, but no new parameters are introduced and the maximization presents no additional difficulty.

Case 2.
The cluster "dies" when all its members have died. This is the "maximum" definition because the lifetime distribution of a cluster of size m is given by the random variable $T_{(m)} = \max_{i=1,\ldots,m} T_i$. In this case the closure property does not extend neatly. The statements in Theorem 12.2.1 concerning η_1 and η_2 still hold, but the term $\xi(z)$ also changes, to

$$\xi^*(z) = \xi(z) \left\{ 1 - \left(1 - e^{-zH_b(t)} \right)^m \right\} \tag{12.5}$$

Both of these definitions can be written as special cases of defining a cluster as "surviving" if at least r of its members are alive. The maximum definition corresponds to $r = 1$ and the minimum definition to $r = m$. The simple closure property applies only to the minimum definition.

Although the failure of the closure property prevents the use of the diagnostic plot with the maximum definition, the derivation of the conditional distribution of frailty among survivors at time t is useful because it makes it possible to employ the E-M algorithm to fit the frailty model, given data only on clusters surviving at time t. This would be applicable in situations where the existence of the cluster can only be recorded if at least one of its members has survived, which represents a form of truncation.

In the E-M approach to estimation, the likelihood is expressed as a function of the values of the frailty Z, which are treated as missing data and are estimated in the E-step.

References

1. Economou, P. and Caroni, C. (2005). Graphical tests for the assumption of Gamma and Inverse Gaussian frailty distributions, *Lifetime Data Analysis*, **11**, 565-582.

2. Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity, *Biometrika* , **71**, 75-83.

3. Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions, *Biometrika*, **73**, 387-396.

4. Shao, J. (1998). *Mathematical Statistics*, Springer-Verlag, NewYork.

5. Vaupel, J.A., Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**, 439-454.

13

On Pseudonormal Extension of the Class of Multivariate Normal Probability Distributions

Jerzy Filus and Lidia Filus

Department of Mathematics and Computer Science, Oakton Community College, Illinois, USA
Mathematics Department, Northeastern Illinois, University, USA

13.1 Introduction

A new class of continuous, easily reversible, transformations R_n onto R_n is defined, and applied in a probabilistic setting. The form of each transformation is basically similar to the affine diagonal mappings on R_n but the parameters depend on the variables in a specific "triangular" way. We call such transformations "pseudoaffine" as, (by analogy to the ordinary affine) they are compositions of "nonsingular pseudolinears" and "pseudotranslations". These transformations, when applied to random vectors of n independent normally distributed random variables, produce, as outputs, random vectors whose joint probability densities have some essential properties similar to those of the n-variate normal. The so obtained new joint probability distributions we call pseudonormal (See Kotz, Balakrishnan, Johnson (2000) pages 217-218.) since 1) the ordinary affine transformations (which are special cases of the pseudoaffines) produce the n-variate normal densities, and 2) the normal densities are special cases of the pseudonormal. Eventually it turns out that the extension of the family of the affine mappings $R_n \to R_n$ corresponds strictly to the extension of the family of the n-variate normals.

The idea of the pseudonormal joint probability distributions of the random vectors, say, $(X_1, \ldots, X_n)$ (as well as of several other pdf classes (first of all pseudoWeibullian) based on the same general pattern, see FF3) has its primary roots in the area of multicomponent system reliability modeling (see, for example, Barlow and Proschan (1975)). In that setting the r. variables $X_1, \ldots, X_n$ are interpreted as stochastically dependent system component life times. In the recent more than four decades the problem of modeling stochastic dependences between life times of system components, by means of their joint probability

distributions, was investigated by many authors, such as Freund (1961), Marshall and Olkin (1967) or Lindley and Singpurwalla (1986) to mention a few among many other excellent positions available. Also see Filus (1991). For a nearly exhaustive references on that subject, see Kotz, Balakrishnan, Johnson (2000). If the class of the pseudonormals is limited to its subclass defined by the pseudotranslations only, other possibility as to characterize the resulting pseudonormals arises. In such cases an alternative definition of these pseudonormals (called "non-multiplicative") is obtained only by adding a nonlinear term or terms to a linear regression function of the corresponding n-variate normal. Such a nonlinear "correction" of the usual linear regression model allows to consider these pseudonormals as, say, nonlinear regression "normal" distributions. In some cases the nonlinear regression normal model may be expected to be more consistent with an empirical data than the ordinary normal. It then appears that applications of the pseudonormals may not be limited to system reliability modeling only. It seems that, sometimes, the pseudonormal models may serve as alternatives to associated normal ones. As we anticipate, in some instances, they might give a better fit to data than their normal counterparts. Within the class of the pseudonormals, symmetric and nonsymmetric densities subclasses are defined. For example, one obtains the symmetric pseudonormal density by adding a cubic term to the normal linear regression function. This procedure indicates a possibility to increase an accuracy of the models, at least in some situations. On the other hand, if a given data shows a slight asymmetric tendency then adding a proper quadratic term might be useful. Generalizing this pattern more, a polynomial, of a reasonable degree, may eventually be applied too.

13.2 The Outline

We recall that by the (ordinary) affine transformation in R_n one usually means composition of a nonsingular linear transformation and a (constant) translation. Let $T_1, \ldots, T_n$ be independent, normally distributed random variables with zero expectation and arbitrary variances $\sigma_1{}^2, \ldots, \sigma n^2$ respectively. Apply to the random vector $(T_1, \ldots, T_n)$ an easily reversible pseudoaffine transformation $(R_n \to R_n)$, given by the following pattern:

$$X_1 = \phi_0 T_1 + \theta_0, \quad X_2 = \phi_1(X_1)T_2 + \theta_1(X_1), \quad \ldots\ldots\ldots\ldots \tag{1}$$

$$X_n = \phi n - 1(X_1, \ldots, X_{n-1})T_n + \phi_{n-1}(X_1, \ldots, X_n - 1) \text{ a.s.},$$

where θ_0 and $\phi_0 \neq 0$ are constants, and $\phi_1, \ldots, \phi_{n-1}, \theta_1, \ldots, \theta_{n-1},\ n = 2, 3, \ldots$ are real continuous functions called "parameter functions". Here we will only consider the case when $\phi_0 = 1$ and $\theta_0 = 0$. We also will use deterministic variables $t_1, \ldots, t_n, x_1, \ldots, x_n$ as real values whenever they are assumed to be

taken on by the random variables $T_1, \ldots, T_n, X_1, \ldots, X_n$ respectively. In this work we are interested in the theory of n-variate probability distributions of the random vectors $(X_1, \ldots, X_n)$, that are outputs in transformations (1), given inputs $(T_1, \ldots, T_n)$ that may only be assumed to be distributed according to the n- variate Gaussian pdfs (so that the marginals independence is not mandatory). The corresponding n-variate probability densities $h(x_1, \ldots, x_n)$ of the transformation outputs will always be given in the factored form as follows:

$$h(x_1, \ldots, x_n) = h_1(x_1)h_2(x_2|x_1) \ldots h_n(x_n|x_1, \ldots, x_n), \tag{2}$$

where, in the case $X_1 = T_1$ a.s., $h_1(x_1)$ is the ordinary $N(0, \sigma_1)$ normal density. For each $j = 2, \ldots, n$ the conditional densities $h_j(x_j|x_1, \ldots, x_{j-1})$ one obtains directly from (1). Notice that, in the above product, each factor $h_j(x_j|x_1, \ldots, x_{j-1})$ is a univariate normal (!) pdf with respect to x_j alone. This fact, as well as the form of the regression function $E[X_j|x_1, \ldots, x_{j-1}] = \theta_{j-1}(x_1, \ldots, x_{j-1})$, that follows, led us to use the name "pseudonormal" (or "exnormal") for the probability densities $h(x_1, \ldots, x_n)$. The conditional standard deviation $s_j(X_j|x_1, \ldots, x_{j-1})$ will simply be expressed as the product $\sigma_j|\phi_{j-1}(x_1, \ldots, x_{j-1})|$.

The usual jacobian of the inverse to (1) turns out simply to be the product:

$$\partial(t_1, \ldots, t_n)/\partial(x_1, \ldots, x_n)| = (|\phi_0||\phi_1(x_1)||\phi_2(x_1, x_2)| \ldots |\phi_{n-1}(x_1, \ldots, x_{n-1})|)^{-1}.$$

Finally, for $j = 2, 3, \ldots, n$, the formula for the conditional densities $h_j(x_j|x_1, \ldots, x_{j-1})$ being factors in (2) is:

$$h_j(x_j|x_1, \ldots, x_{j-1}) = (\sigma_j|\phi_{j-1}(x_1, \ldots, x_{j-1})|\sqrt{2\pi})^{-1} \times$$

$$\exp[-(1/2\sigma_j^2(\phi_{j-1}(x_1, \ldots, x_{j-1}))^2)(x_j - \theta_{j-1}(x_1, \ldots, x_{j-1}))^2] \tag{3}$$

and therefore the joint pseudonormal pdfs are entirely determined in the form (2).

13.3 On some representative Bivariate pseudonormals

In FF2 we considered the 2-dimensional transformations "from independent normals to pseudonormals", which were subjected to the following scheme:

$$X_1 = T_1, \quad X_2 = \phi(X_1 - \mu_1)T_2 + a(X_1 - \mu_1) + \theta^{\star\star}(X_1 - \mu_1), \tag{4}$$

where T_1, T_2 were assumed to be independent r. variables with $N(0, \sigma_1), N(0, \sigma_2)$ densities respectively. The parameter function $\theta^{\star}(T_1 - \mu_1)$ was referred to as a nonlinear term of the regression function $E[X_2|x_1]$. Now realize that the

transformation (4) may be regarded as a composition of the two following transformations: 1) the affine transformation:

$$T_1^\star = T_1, \quad T_2^\star = T_2 + a(X_1 - \mu_1), \tag{5}$$

that transforms independent normals T_1, T_2 to the dependent normal $(T_1^\star, T_2^\star)$, and 2) the pseudoaffine one:

$$X_1 = T_1^\star, \quad X_2 = \phi(X_1 - \mu_1)T_2^\star + \theta^\star(X_1 - \mu_1), \tag{6}$$

which transforms the (arbitrary) normal r. vector $(T_1^\star, T_2^\star)$, with a correlation coefficient $\rho = (a\sigma_1)/\sigma_2$, into the actual pseudonormal r. vector (X_1, X_2) being considered. Such a procedure dramatically simplifies calculations in comparison with those performed in FF2. We refer to that paper, and, in order to unify the notation, we will use back the symbol (T_1, T_2) instead of the above $(T_1^\star, T_2^\star)$ to denote the input random variables in (6). In this way we defined the transformation (6) (instead of the more complicated (4) used in FF2) as defined on arbitrary Gaussian random vectors ((T_1, T_2) with a correlation coefficient ρ as above. The simplification gained in formula (6) in comparison to (4) turns out to give a significant improvement in calculation efficiency regardless a cost of some (mild) complexity now associated with the dependence of the input random variables (T_1, T_2). Also we replace back the symbol $\theta^\star()$ in (6) by $\theta()$ for the nonlinear term. Notice that the conditions $E[T_1] = E[T_2] = 0$ hold. With all the new meanings we rewrite (6) as:

$$X_1 = T_1, \quad X_2 = \phi(X_1)T_2 + \theta(X_1). \tag{6*}$$

The regression function of the pseudonormal r. vector (X_1, X_2) defined by (6*) now is: $E[X_2|x_1] = ax_1 + \theta(x_1)$, with $a = \rho\sigma_2/\sigma_1$. In FF2 it is shown that for symmetric pdfs of (X_1, X_2) we have $E[X_2] = 0$, whenever $E[T_2] = 0$. The nonlinear part $\theta(x_1)$ of the regression function enriches the original Gaussian stochastic dependence structure of (T_1, T_2) determined by ρ. The problem of finding the r$^{\text{th}}$ moments for the marginal variable X_2, $r = 2, 3, \ldots$, now simplifies as only two terms in (6*) are to be raised to the r$^{\text{th}}$ power instead of three in the expression (4) as it was the case in FF2.

13.4 On Some Bivariate Analytic Examples

Example 1: The Symmetric Case. In the following example the bivariate density function $h(x_1, x_2)$ is assumed to be symmetric in the sense that $h(-x_1, -x_2) = h(x_1, x_2)$ and thus $h_2(-x_2) = h_2(x_2)$, where $h_2(x_2)$ denotes the marginal density of the r. variable X_2. Recall that in such a case we have $E[X_2] = 0$ whenever $E[T_2] = 0$. For more details about the symmetry of pseudonormals see FF2. Consider the following pseudoaffine (or rather

pseudotranslation) transformation:

$$X1 = T1, \quad X_2 = T_2 + AX_1^{(2k+1)} \tag{7},$$

where k is a positive integer, and (T_1, T_2) is an arbitrary random vector with the joint density $f(t_1, t_2) = f_1(t_1)f(t_2|t_1) =$

$$(1/(\sigma_1\sqrt{2\pi}))exp[-t_1^2/(2\sigma_1^2)](1/\sigma_2\sqrt{2\pi(1-\rho^2)})exp[-(t_2 - at_1)2/(2\sigma_2^2(1-\rho^2))], \tag{8}$$

where ρ is the ordinary correctional coefficient of (T_1, T_2) and $\rho = a\sigma_1/\sigma_2$. Using (7) and (8) one obtains the joint pdf of the random vector (X_1, X_2) in the form: $h(x_1, x_2) = h_1(x_1)h_2(x_2|x_1)$, where $h_1(x_1) = f_1(x_1)$ is the same as the first factor in (8) upon replacing t_1 by x_1 , and

$$h_2(x_2|x_1) = (1/\sigma_2\sqrt{2\phi(1-\rho^2)})exp[-(x_2 - ax_1 - Ax_1^{(2k+1)})^2/2\sigma_2^2(1-\rho^2)].$$

It is clear, especially when the parameter $'A'$ is small in comparison to $'a'$, that the new bivariate pseudonormal density $h(x_1, x_2)$ arises as the result of adding the nonlinear correcting term $Ax_1^{(2k+1)}$ to the original linear regression function $E[T_2|x_1] = ax_1$ of the bivariate normal (T_1, T_2) , when $T_1 = x_1$. This case is specially interesting when $k = 1$. We call it the cubic correction. For the similar quadratic correction see **Example 2**, for $k = 1$.

Evaluating the moments and other parameters of the pdf considered in this example is much easier than in Example 1S in FF2. After elementary calculations one obtains the expectations $E[X_1] = 0, E[X_2] = 0$, and $E[X_2|x_1] = ax_1 + Ax_1^{(2k+1)}$. Also one obtains the variance of X_2 in the form

$$Var(X_2) = s_2^2 = \sigma_2^2 + A^2 j(4k+1)\sigma_1^{(4k+2)}, \tag{9}$$

where $j(w) = (1)(3)(5)\dots(w)$ for any positive odd integer w, so that $j(4k+1)\sigma_1^{(4k+2)}$ is the $(4k+2)^{\text{th}}$ central moment of X_1 (Recall that the r.v's X_1 and T_1 have the same normal pdf). The covariance and the correlation coefficient c of (X_1, X_2) are

$$E[X_1X_2] = \int_{-\infty}^{\infty} x_2h_2(x_2|x_1)dx_2x_1h_1(x_1)dx_1 = as_1^2 + Aj(2k+1)\sigma_1^{(2k+2)},$$

and $c = \rho + j(2k+1)A(s_1^{(2k+1)}/s_2)$ respectively, with $s_1 = \sigma_1$, and s_2 given by (9). Notice that $s_2 \to \sigma_2$, $c \to \rho$ as $A \to 0$, where $\rho = as_1/s_2$ is the original correlation coefficient of the normal r. vector (T_1, T_2).

Because of space limitation next two Examples of Bivariate nonSymmetric Pseudonormals will only be sketched as described by the following defining pseudoaffine transformations:

Example 2. A class of nonmultiplicative pseudonormal pdfs, which seems to be quite interesting with respect to possible applications, is given by the following class of the pseudotranslations: $X_1 = T_1,\ X_2 = T_2 + AX_1^{2k}$, where k is a positive integer, and the input r. vectors (T_1, T_2) are arbitrary normal with any correlation coefficients ρ. For more details concerning this case see FF4, where, between others, the central moments up to the fourth are obtained in a relatively simple analytic form.

Example 3. A class of the multiplicative densities can be given by the following class of pseudolinear transformations, applied to the same as above normal r. vectors $(T_1, T_2) :\ X_1 = T_1,\ X_2 = A\cosh(\lambda X_1)T_2$, with A and λ being positive real constants.

13.5 Final Remarks

We want to make a note about other possible applications of the pseudolinear transformations, which one obtains from (1) by letting $\theta_0 = \theta_1(x_1) = \ldots = \theta_{n-1}(x_1, \ldots, x_{n-1}) = 0$.

1. In system reliability modeling we assume that the input random variables $T_1, \ldots, T_n$ are independent exponentials with the expectations $\alpha_1, \ldots, \alpha_n$ respectively rather than the normals. As a result one obtains a class of joint probability densities $h(x_1, \ldots, x_n)$ of the output $(X_1, \ldots, X_n)$ in the form (2), where $h_1(x_1)$ is the ordinary exponential with an expectation α_1, and for each $j = 2, \ldots, n$ the conditional exponential densities are:

$$h_j(x_j|x_1, \ldots, x_{j-1}) = (1/\alpha_j|\phi_{j-1}(x_1, \ldots, x_{j-1})|)exp[-x_j/\alpha_j|\phi_{j-1}(x_1, \ldots, x_{j-1})|]. \tag{10}$$

2. Let, as before, $T_1, \ldots, T_n$ be independent exponential random variables. Extend the previously defined pseudolinear transformations pattern into the pseudopower: $X_1 = T_1^{\gamma_1},\ X_2 = \phi_1(X_1)T_2^{\gamma_2},\ \ldots,\ X_n = \phi_{n-1}(X_1, \ldots, X_{n-1})T_n^{\gamma_n}$ with arbitrary positive reals $\gamma_i = 1/\beta_i$, for $i = 1, \ldots, n$. Next gain in generality can be obtained by admitting (for $i = 2, 3, \ldots, n$) β_i to depend on the r. variables $X_1, \ldots, X_{i-1}$, while β_1 remains constant. As a result one obtains the joint density $h(x_1, \ldots, x_n)$ of the r. vector $(X_1, \ldots, X_n)$ in a form similar to (10), however the conditional densities $h_j(x_j|x_1, \ldots, x_{j-1})$ now are Weibullian, each with respect to x_j alone. For $j = 1, 2, \ldots, n$ the shape parameters are β_j's respectively. For that reason the density $h(x_1, \ldots, x_n)$ is called "pseudoweibullian". It can be shown that in the foregoing case one can weaken the assumption that the variables $T_1, \ldots, T_n$ are exponential, and allow them to be arbitrary independent Weibullians. For more details on the reliability applications of the so defined pseudoexponential and pseudoweibullian models see FF3.

3. In addition to the context of this paper, some nice theoretical results obtained are to be mentioned. They mainly concern an invariance of the classes of

the pseudonormals and the pseudoexponentials with respect to the pseudoaffine, as well as the pseudoWeibullians with respect to pseudopower transformations. Speaking briefly, any pseudonormal (so also a normal) or pseudoexponential input, in any pseudoaffine transformation, results in a pseudonormal or a pseudoexponential output respectively. The same can be said about transforming pseudoWeibullians through the pseudopowers so, in particular, through the pseudoaffines. The proofs and some analysis of these properties can be found in FF1.

References

1. Barlow, R.E. and Proshan, F. (1975). Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Windston, New York.

2. Filus, J.K. (1991). On a type of dependencies between Weibull life times of system components, Reliability Engineering and System Safety, Vol. 31, No. 3, 267-280.

3. **FF1**: Filus, J.K. and Filus, L.Z. (2000). A class of generalized multivariate normal densities, Pakistan J. of Statist., Vol. 16 (1), pp 11-32.

4. **FF2**: Filus, J.K. and Filus, L.Z. (2001). On some bivariate pseudonormal densities, Pakistan J. of Statist., 1, Vol 17, 1-19.

5. **FF3**: Filus, J.K. and Filus, L. Z. (2006). On Some New Classes of Multivariate Probability Distributions. Pakistan J. of Statist., 1. Vol 22, pp 21- 42.

6. **FF4**: Filus, J.K. and Filus, L. Z., (2003). On Two New Methods for Constructing Multivariate Probability Distributions with System Reliability Applications. Technical Report, No. 03-01-28 Department of Mathematics, Northeastern Illinois University, Chicago, IL. 60625 USA.

7. Freund, J. E. (1961). A bivariate extension of the exponential distribution, J. Amer. Statist. Assoc., Vol. 56, 971- 977.

8. Kotz, S., Balakrishnan, N., Johnson, N. L., (2000). Continuous Multivariate Distributions. Volume 1. Second Edition. J. Wiley & Sons, Inc, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto.

9. Lindley, D.V. and Singpurwalla, N.D. (1986). Multivariate distributions for the life lengths of components of a system sharing a common environment, J. Appl. Prob.23, 418-431

10. Marshall, A.W. and Olkin, I. (1967). A generalized bivariate exponential distribution, J. of Appl. Prob. 4, 291-302.

14

On Virtual Age of Degrading Systems

Maxim Finkelstein

Department of Mathematical Statistics, University of the Free State,
PO Box 339, 9300 Bloemfontein, Republic of South Africa,
(e-mail: FinkelM@sci.uovs.ac.za)
and
Max Planck Institute for Demographic Research, Rostock, Germany

Abstract: Two approaches to defining virtual age of a degrading system are considered. The first one is based on the fact that deterioration depends on environment. In a more severe environment deterioration is more intensive, which means that objects are aging faster and therefore, the corresponding virtual age is larger than the calendar age in a baseline environment. The second approach is based on considering an observed level of individual degradation and comparing it with some average, 'population degradation'.

Keywords and phrases: Virtual age, Degradation, Aging distributions, Failure rate, Mean remaining lifetime

14.1 Introduction

Lifetimes of degrading (deteriorating) systems can be effectively modeled by aging distributions. The simplest and probably the most natural is the class of distributions with increasing failure rates (IFR). It is clear that an age of a system, as some overall trivial marker of deterioration, is really informative only for degrading objects. This age is the same for all individuals in a population, which are simultaneously incepted into operation. We shall call this chronological age the *statistical age*.

Deterioration usually depends on environment. Deterioration under a more severe environment is more intensive, which means that objects are aging faster. Therefore, we discuss a *statistical virtual age*, which is defined for degradation comparison under different environments (stresses). We also introduce an *information-based virtual age* of a system. If, for instance, an individual of 50 years old looks like and has vital characteristics (blood pressure, level of cholesterol etc) are as of a 30 years old one, we can say that this observation indicates that his virtual age could be 30.

Another challenging problem to be considered is to define the virtual age of a system with components in series having different virtual ages.

14.2 Virtual Age In Repairable Systems

We start with some introductory, helpful considerations on a notion of a virtual age for repairable systems (Kijima (1988), Finkelstein (2000)).

A convenient mathematical description of repair processes uses a concept of stochastic (or failure) intensity λ_t (Aven and Jensen (1999)). Consider, for example, a renewal process (perfect instantaneous repair) with underlying distribution $F(t)$ and the failure rate $\lambda(t)$. Then

$$\lambda_t = \sum_{n=0}^{\infty} \lambda(t - T_n) I(T_n \leq t < T_{n+1}). \tag{14.1}$$

Denote by A_t the *age process*, which corresponds to the renewal process (14.1):

$$A_t = \sum_{n=0}^{\infty} (t - T_n) I(T_n \leq t < T_{n+1}). \tag{14.2}$$

Thus, this stochastic process starts at $t = 0$ as a linear function with a unit slope. It jumps again to 0 at T_1, the time of the first repair, etc. The age of a repairable system in this case is just time elapsed since the last repair. Note, that as a minimal repair (note a perfect one!) does not change the age of a system, the corresponding age process is deterministic: $A_t = t$.

Consider now intermediate between the perfect and minimal level of repair, which brings an important notion of a *virtual age*. Assume that repair at $t = t_1$ decreases the age of a system not to 0 as for a perfect repair, but to $\nu_1 = qt_1$, $0 < q < 1$, and the system starts the second cycle with this initial age in accordance with the Cdf $1 - \bar{F}(t_1 + t)/\bar{F}(t_1)$. This age is called the virtual age. $F(t)$ is assumed to be IFR in this approach.

14.3 Statistical Virtual Age

Consider a degrading system in a fixed baseline environment with the Cdf of time to failure $F_b(t)$. The chronological age t will be called the *statistical age*. Let another statistically identical system be operating in a more severe environment with the Cdf of time to failure $F_s(t)$. We want to establish an age correspondence between these two regimes. Degradation under the second regime is more intensive therefore the time to reach the same level of degradation, as under the baseline one, will be smaller. We shall call the corresponding time the *statistical virtual age* of the second system.

We will describe this definition in mathematical terms for a specific model. Assume that the lifetimes for two regimes are ordered in the sense of a usual stochastic ordering:

$$\bar{F}_s(t) < \bar{F}_b(t), \quad t \in [0, \infty), \tag{14.3}$$

Inequality (14.3) implies the following equation:

$$F_s(t) = F_b(W(t)), \quad W(0) = 0,\ W(t) > t,\ t \in (0, \infty), \tag{14.4}$$

which can be interpreted as a generalized Accelerated Life Model (ALM) (Cox and Oakes (1984)) with a scale transformation function $W(t)$, which in its turn can be interpreted as an additive degradation function $W(t) = \int_0^\infty w(u)du$, where $w(t)$ has a meaning of a speed of degradation. Therefore, the statistical virtual age of the second system is $W(t)$, compared with the statistical age t of a system in a baseline environment.

When the failure rates are given, or estimated from the data, relation (14.4) can be viewed as an equation for obtaining the statistical virtual age $W(t)$:

$$\int_0^t \lambda_s(u)du = \int_0^{W(t)} \lambda_b(u)du. \tag{14.5}$$

14.4 Information-based Virtual Age

In the previous section a system was considered as a black box. Observation of a state of a system at time t can give an indication (under certain assumptions) of its age, defined by the level of deterioration.

We start with a meaningful reliability example when the number of observed operable components defines the corresponding level of deterioration.

Example 1. Consider a system of $n+1$ components (one initial component and n cold standby identical ones) with constant failure rates λ. Denote the system's lifetime random variable by T_{n+1}. The corresponding Cdf is

$$F_{n+1}(t) \equiv Pr[T_{n+1} \le t] = 1 - e^{-\lambda t} \sum_0^n \frac{(\lambda t)^i}{i!} \tag{14.6}$$

with an increasing failure rate.

In order to obtain the corresponding information-based virtual age to be compared with the statistical age t, consider, firstly, the following conditional expectation:

$$\begin{aligned} D(t) &\equiv E[N(t) \,|\, N(t) \le n] = E[N(t) \,|\, T_{n+1} > t] \\ &= \frac{e^{-\lambda t} \sum_0^n i \frac{(\lambda t)^i}{i!}}{e^{-\lambda t} \sum_0^n \frac{(\lambda t)^i}{i!}}, \end{aligned}$$

where $N(t)$ is the number of events in the interval $[0,t]$ for the Poisson process with rate λ. The function $D(t)$ is monotonically increasing, $D(0) = 0$ and $\lim_{t\to\infty} = n$. This function defines the 'mean degradation curve' for the system.

Denote the information-based virtual age by $V(t)$. Our *definition* is:

$$V(t) = D^{-1}(k). \tag{14.7}$$

If $k = D(t)$, then: $V(t) = D^{-1}(D(t)) = t$. Similar:

$$k < D(t) \quad \Rightarrow \quad V(t) < t, \qquad k > D(t) \quad \Rightarrow \quad V(t) > t.$$

The general case of degrading objects can be considered in the same line. Let D_t be an increasing, smoothly varying (predictable) stochastic process of degradation with a mean $D(t)$, which defines the statistical age of our object as t. We also assume for simplicity that this is a process with independent increments. Then d_t - observation at time t *defines* the information-based virtual age as $V(t) = D^{-1}(d_t)$.

Alternatively, $V(t)$ can be defined via the information-based remaining lifetime (Finkelstein, 2001). The statistical (conventional) mean remaining lifetime (MRL) at t of a system with the Cdf $F(x)$ is defined in a standard way as:

$$M(t) = \int_0^\infty F(x|t)dx = \int_0^\infty \frac{\bar{F}(t+x)}{\bar{F}(t)}dx \tag{14.8}$$

and we must compare it with the mean information-based remaining lifetime, denoted by $M_I(t)$. *Define* the information-based virtual age in this case as

$$V(t) = t + (M(t) - M_I(t)). \tag{14.9}$$

Note that the idea of our definition (14.9) is in adding (subtracting) to the chronological age t the gain (loss) in the remaining lifetime due to additional information.

Example 2. Consider a system of 2 i.i.d components in parallel with exponential Cdfs. Then $\bar{F}(t) = e^{-2\lambda t} - 2e^{-\lambda t}$ and

$$M(t) = \int_0^\infty \frac{2e^{-\lambda t} - e^{-2\lambda t}e^{-\lambda x}}{2 - e^{-\lambda x}}dx.$$

Assume that our observation at time T is: two operable components. Then $M_I(t) = 1.5\lambda$ and $0 < V(t) < t$. If observation is one operable component, then the virtual age is larger than t: $V(t) > t$.

14.5 Virtual Age of a Series System

In this section possible approaches to defining a virtual age of a series system of degrading components with different virtual ages will be considered. In a conventional setting all components have the same chronological age and therefore this problem does not exist. However, it is really important in different applications (specifically, biological) to obtain a virtual age of a series system.

We start with considering the statistical virtual age discussed in Section 14.3. The survival functions of a series system of statistically independent components under the baseline and a more severe environment are

$$\bar{F}_b(t) = \prod_1^n \bar{F}_{bi}(t); \qquad \bar{F}_s(t) = \prod_1^n \bar{F}_{bi}(W_i(t)), \tag{14.10}$$

respectively, where $W_i(t)$ is the scale transformation function for the ith component and we assume that the model (14.4) holds for every component. Thus, each component has its own statistical virtual age $V_i(t) = W_i(t)$, whereas the virtual age for the system $V(t) = W(t)$ can be obtained from the following equation:

$$\int_0^{W(t)} \sum_1^n \lambda_{bi}(u)du = \sum_1^n \int_o^{W_i(t)} \lambda_{bi}(u)du. \tag{14.11}$$

Going back to the information-based virtual age, as the first choice, we shall weight ages in the series system of n degrading components in accordance with the importance of the components with respect to the failure of the system. Let $V_i(t)$ denote the information-based virtual age of the ith component with a failure rate $\lambda_i(t)$ in a series system of n statistically independent components. Then the virtual age of a system at time t *is defined* as an expected value of the virtual age of a failed in $[t, t + dt)$ component:

$$V(t) = \sum_1^n \frac{\lambda_i(t)}{\lambda_s(t)} V_i(t),$$

where $\lambda_s(t) = \sum_1^n \lambda_i(t)$ is the failure rate of the series system.

The second approach is based on the notion of the MRL function. Thus,

$$\bar{F}(x) = \prod_1^n \bar{F}_i(x), \qquad \bar{F}(x|t) = \frac{\bar{F}(x+t)}{\bar{F}(x)} = \prod_1^n \bar{F}_i(x|t).$$

Denote now by $F_{I,i}(x,t)$ the information-based Cdf of the remaining lifetime for the ith component. Then

$$M(t) = \int_0^\infty \prod_1^n \bar{F}_i(x|t)dx, \qquad M_I(t) = \int_0^\infty \prod_1^n \bar{F}_{I,i}(x,t)dx$$

and eventually, equation (14.9) should be used for obtaining the corresponding information-based virtual age of a series system in this case.

References

1. Aven, T. and Jensen, U. (1999). *Stochastic Models in Reliability.*, Springer.

2. Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London.

3. Finkelstein, M.S. (2000). Modeling a process of non-ideal repair, In *Recent Advances in Reliability Theory.* (Eds., Limnios N., Nikulin M.), pp. 41-53, Birkhauser.

4. Finkelstein, M.S. (2001). How does our n-component system perform? *IEEE Transactions on Reliability*, **50**, 414-418.

5. Kijima, M. (1989) Some results for repairable systems with general repair. *J. Appl. Prob.* **26**, 89-102.

15

Constant–sum Models and Interval–censored Data

Guadalupe Gómez[1], Ramon Oller[2] and M. Luz Calle[2]

[1] *Universitat Politècnica de Catalunya*
[2] *Universitat de Vic*

Abstract: In survival data analysis the interval censoring problem has been usually treated via maximum likelihood inferences. Standard methods suppose that conditions producing censoring do not affect the survival process in order to justify the use of a simpler expression of the likelihood function. This paper is about formal conditions to ensure the validity of such a simplified likelihood.

Keywords and phrases: Constant-sum condition, Identifiability, Interval-censored data, Noninformative condition

15.1 Introduction

Interval censoring mechanisms arise when the event of interest cannot be directly observed and it is only known to have occurred during a random interval of time. This type of censored data has been extensively analyzed during the last years. Inference methods are mainly based on the simplified likelihood we would obtain if the censoring intervals were fixed in advance, ignoring the randomness of the intervals. Such likelihood–based inferences are correct when the censoring process does not affect the lifetime variable. Moreover, there are situations where this noninformative assumption does not hold but the use of the simplified likelihood is still correct. In this work, we talk about these informative situations. We introduce and discuss the results given in Oller *et al.* (2004).

Let T be the positive random variable of interest representing the time until the occurrence of a certain event $\mathcal{E}$ with unknown right-continuous distribution function $W(t) = \text{Prob}\{T \leq t\}$ and support $\mathcal{D}_W = \{t > 0 : dW(t) > 0\}$. Data is said to be interval–censored when the time to $\mathcal{E}$ is unknown and instead of this time we observe a time interval $\lfloor L, R \rfloor$ where L is the last observed time before the event $\mathcal{E}$ has occurred and R indicates the first time the event $\mathcal{E}$ has

been observed. We use the $\lfloor L, R \rfloor$ notation to indicate an interval that can be closed, open or half open depending on the censoring model. We are in fact formally observing a random censor vector (L, R), such that $T \in \lfloor L, R \rfloor$ with probability 1. Thus, a model for interval censored data is described by a joint distribution $F_{L,R,T}$ with range $\big\{(t, l, r) : 0 \leq l \leq t \leq r < 1\big\}$.

Based on the marginal law of the observables $F_{L,R}$ and a sample of n observations $\lfloor l_1, r_1 \rfloor, \ldots, \lfloor l_n, r_n \rfloor$, the full likelihood can be expressed as,

$$L_0 = \prod_{i=1}^{n} dF_{L,R}(l_i, r_i) = \prod_{i=1}^{n} P(L \in dl_i,\, R \in dr_i,\, l_i \leq T \leq r_i) \tag{15.1}$$

If the observed intervals are treated as fixed in advance and we ignore their randomness, then the likelihood simplifies as,

$$L = \prod_{i=1}^{n} P_W(\lfloor l_i, r_i \rfloor) = \prod_{i=1}^{n} P(l_i \leq T \leq r_i) \tag{15.2}$$

In the next section, we define conditions under which the nonparametric maximum likelihood estimator (NPMLE) of the lifetime distribution can be based on this simplified likelihood (15.2).

15.2 Noninformative and constant–sum models

In studies where interval–censored data arise because individuals are intermittently inspected, it is usually assumed that the inspection process is independent of T. This independence written in terms of (L, R) and T reduces to the following noninformative condition.

Definition 15.2.1 *A model $F_{L,R,T}$ is noninformative if the following condition holds:*

$$dF_{L,R|T}(l, r|t) = \frac{dF_{L,R}(l, r)}{P_W(\lfloor l, r \rfloor)}\ \mathbf{1}_{\{t \in \lfloor l, r \rfloor\}}(l, r). \tag{15.3}$$

This property has been introduced in the papers of Self and Grossman (1986) and Gómez *et al.* (2004). In a more general censoring framework, Heitjan and Rubin (1991) and Gill *et al.* (1997) develop and characterize the analogous notion of coarsening at random conditions. In Oller *et al.* (2004) different characterizations for the noninformative condition are given and their equivalence is shown. They also introduce a weaker condition, namely the constant–sum condition, which is sufficient for the validity of the simplified likelihood (15.2) in a nonparametric estimation of the lifetime probability distribution W. The constant–sum condition for interval censoring is an extension of the same notion in Williams and Lagakos (1977), in the context of right censoring, and Betensky (2000), in the context of current status data.

Definition 15.2.2 *A model $F_{L,R,T}$ is constant–sum if the following condition holds:*

$$\iint_{\{(l,r):t\in\lfloor l,r\rfloor\}} \frac{dF_{L,R}(l,r)}{P_W(\lfloor l,r\rfloor)} = 1 \qquad \forall t \in \mathcal{D}_W \tag{15.4}$$

Theorem 15.2.1 *If a censoring model $F_{L,R,T}$ is constant–sum, the NPMLE of the lifetime distribution function can be obtained through maximization of the simplified likelihood (15.2) .*

Clearly, if a censoring model is noninformative, then the model is constant-sum. The reciprocal is not true. However, as it is showed in Oller *et al.* (2004), if a censoring model F_{T_1,L_1,R_1} is constant–sum, then there always exists a noninformative model, F_{T_2,L_2,R_2}, with the same marginal distributions, $W_2 = W_1$ and $F_{L_2,R_2} = F_{L_1,R_1}$.

Further discussion about the relationship between the noninformative and the constant–sum conditions is given by Lawless (2004). The author consider situations where an inspection process defines the censoring observations. When the inspection process depends on T, Lawless (2004) proves that the constant–sum property is equivalent to the existence of an alternative inspection process which is independent of T and which gives the same distribution for the observables, $F_{L,R}$, as the underlying true inspection process.

15.3 Illustration

Here we present an example of constant–sum model which illustrates previous results. Let $\mathcal{D}_W = \{0,1,2,3,4\}$ be the support of the lifetime variable and $\mathcal{D}_{F_{L,R}} = \{[0,1],[0,2],[2,4],[3,3]\}$ the observable censoring intervals. We consider the model determined by the joint probability between the lifetime variable and the observables, $dF_{T,L,R}(t,l,r)$, given by Table 15.1.

Table 15.1: *Joint probability $dF_{T,L,R}$ of a constant–sum model.*

$\lfloor l,r\rfloor$ / t	[0,1]	[0,2]	[2,4]	[3,4]	$dW(t)$
0	1/24	3/24	0	0	1/6
1	3/24	1/24	0	0	1/6
2	0	1/6	1/6	0	1/3
3	0	0	1/24	3/24	1/6
4	0	0	3/24	1/24	1/6
$dF_{L,R}(l,r)$	1/6	1/3	1/3	1/6	1

It is easy to verify that this model holds the constant–sum condition (15.4) for each $t \in \{0,1,2,3,4\}$. For instance, for $t = 1$ the constant–sum condition is

$$\sum_{\{(l,r):1\in\lfloor l,r\rfloor\}} \frac{dF_{L,R}(l,r)}{P_W(\lfloor l,r\rfloor)} = \frac{dF_{L,R}(0,1)}{P_W(\lfloor 0,1\rfloor)} + \frac{dF_{L,R}(0,2)}{P_W(\lfloor 0,2\rfloor)} = \frac{\frac{1}{6}}{\frac{1}{6}+\frac{1}{6}} + \frac{\frac{1}{3}}{\frac{1}{6}+\frac{1}{6}+\frac{1}{3}} = 1.$$

However, this model does not hold the noninformative condition. For instance, $dF_{L,R|T}(0,2|0) = 3/4$, while $dF_{L,R}(0,2)/P_W([0,2]) = 1/2$, so condition (15.3) fails.

Henceforth, we illustrate the result in Theorem 15.2.1 by considering a sample of n_1 individuals with observed intervals $[0,1]$, n_2 individuals with $[0,2]$, n_3 with $[2,4]$ and n_4 with $[3,4]$. In this case, the likelihood function (15.1) is given by

$$L_0 = [dF_{L,R}(0,1)]^{n_1}\,[dF_{L,R}(0,2)]^{n_2}\,[dF_{L,R}(2,4)]^{n_3}\,[dF_{L,R}(3,4)]^{n_4}\,.$$

Using the following factorization

$$dF_{L,R}(l,r) = P_W(\lfloor l,r\rfloor)\cdot\frac{dF_{L,R}(l,r)}{P_W(\lfloor l,r\rfloor)}$$

and parametrizing as $\theta_t = P_W(t)$ for $t = 0,1,2,3,4$ and

$$\gamma_j = dF_{L,R}(l_j,r_j)/P_W(\lfloor l_j,r_j\rfloor) \quad \text{for } (l_j,r_j) = (0,1),(0,2),(2,4),(3,4)$$

we have,

$$L_0 = [(\theta_0+\theta_1)\gamma_1]^{n_1}[(\theta_0+\theta_1+\theta_2)\gamma_2]^{n_2}[(\theta_2+\theta_3+\theta_4)\gamma_3]^{n_3}[(\theta_3+\theta_4)\gamma_4]^{n_4}\,.$$

The MLE of the parameters θ and γ is obtained by maximizing the logarithm of the likelihood, $\log(L_0)$, subject to the constraints:

(i) $0 \le \theta_t \le 1$

(ii) $\theta_0+\theta_1+\theta_2+\theta_3+\theta_4 = 1$

(iii) $0 \le \gamma_j \le 1$

(iv) $(\theta_0+\theta_1)\gamma_1 + (\theta_0+\theta_1+\theta_2)\gamma_2 + (\theta_2+\theta_3+\theta_4)\gamma_3 + (\theta_3+\theta_4)\gamma_4 = 1.$

Constraint (iv) is needed because $dF_{L,R}(0,1)+dF_{L,R}(0,2)+dF_{L,R}(2,4)+dF_{L,R}(3,4) = 1$, and it can be equivalently written as

(iv) $\theta_0(\gamma_1+\gamma_2) + \theta_1(\gamma_1+\gamma_2) + \theta_2(\gamma_2+\gamma_3) + \theta_3(\gamma_3+\gamma_4) + \theta_4(\gamma_3+\gamma_4) = 1.$

Since the model is constant-sum, in the maximization process we have to add equation (15.4) as a new constraint, that is,

(v) $\gamma_1 + \gamma_2 = 1$, $\gamma_2 + \gamma_3 = 1$ and $\gamma_3 + \gamma_4 = 1$.

However condition (iv) can be derived from conditions (v) and (ii): $\theta_0(\gamma_1+\gamma_2)+\theta_1(\gamma_1+\gamma_2)+\theta_2(\gamma_2+\gamma_3)+\theta_3(\gamma_3+\gamma_4)+\theta_4(\gamma_3+\gamma_4) = \theta_0+\theta_1+\theta_2+\theta_3+\theta_4 = 1$, and hence condition (iv) could be omitted in the maximization process and it is no longer a constraint between the parameters θ and γ. Therefore, the MLE of θ can be obtained maximizing the simplified likelihood

$$\log(L) = n_1 \log(\theta_0+\theta_1)+n_2 \log(\theta_0+\theta_1+\theta_2)+n_3 \log(\theta_2+\theta_3+\theta_4)+n_4 \log(\theta_3+\theta_4)$$

under the constraints (i) and (ii).

15.4 Identifiability of the lifetime distribution

This section is devoted to study the identifiability of the lifetime distribution W on the basis of the assumed support of the lifetimes $\mathcal{D}_W$ and the distribution for the observables $F_{L,R}$. We assume a known lifetime support, $\mathcal{D}_W$, which is not necessarily equal to the usual assumption $\mathcal{D}_W = (0, \infty)$.

Definition 15.4.1 *Given a censoring model $F_{T,L,R}$, we say that W is nonidentifiable when there exists a censoring model having different lifetime distribution but sharing the same lifetime support $\mathcal{D}_W$ and the same distribution for the observables $F_{L,R}$.*

Generally, W will not be identifiable unless we assume some kind of restriction on the model. In fact, the following theorem shows that if we are restricted to the class of constant–sum models, the probabilities assigned by the lifetime distribution to the observable intervals $\lfloor l, r \rfloor$ can be identified from the pair $(\mathcal{D}_W, F_{L,R})$. To assure the complete identifiability of W, however, additional conditions on the observables support will be necessary.

Theorem 15.4.1 *Let $F_{T,L,R}$ and F_{T^*,L^*,R^*} be constant–sum models such that $(\mathcal{D}_W, F_{L,R}) = (\mathcal{D}_{W^*}, F_{L^*,R^*})$, then $P_W(\lfloor l, r \rfloor) = P_{W^*}(\lfloor l, r \rfloor)$ $dF_{L,R}$–almost surely.*

There are specific situations where it is possible to ensure complete identifiability. For instance, when uncensored data are allowed for the whole support of the lifetime variable, that is, when $dF_{L,R}(t,t) > 0$ for any $t \in \mathcal{D}_W$. This identifiability assumption is rather mild and it is typically satisfied in right censored data and doubly censored data applications.

When $\mathcal{D}_W = (0, \infty)$ and every observable arises from a random inspection process with discrete support (L and R lie in a set $\{a_0, a_1, \ldots, a_k\}$ with $0 = a_0 < a_1 < \cdots < a_k = +\infty$), we can show that the probabilities $P_W((a_{j-1}, a_j])$

are identifiable if and only if every value in the set $\{a_0, a_1, \ldots, a_k\}$ is being inspected. When the support of the inspection times is not finite, the constant–sum property is not enough to assure complete identifiability, we need a countable number of inspections for each individual and independence between the inspection process and the lifetime variable instead. In that case, we can also show that if the support of L or R covers $\mathcal{D}_W = (0, +\infty)$, then W is completely identifiable.

References

1. Betensky, R. A. (2000). On nonidentifiability and noninformative censoring for current status data, *Biometrika*, **87**, 218–221.

2. Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). Coarsening at random: characterizations, conjectures, counter–examples. In *Proceedings First Seattle Symposium on Biostatistics: Survival Analysis*, pp. 255–294, Springer–Verlag.

3. Gómez, G., Calle, M. L. and Oller, R. (2004). Frequentist and bayesian approaches for interval–censored data, *Statistical Papers*, **45**, 139–173.

4. Heitjan, D. F. and Rubin D. B. (1991). Ignorability and coarse data, *The Annals of Statistics*, **19**, 2244–2253.

5. Lawless, J. F. (2004). A note on interval–censored lifetime data and the constant–sum condition of Oller, Gómez and Calle (2004), *The Canadian Journal of Statistics*, **32**, 327–331.

6. Oller, R., Gómez, G. and Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood, *The Canadian Journal of Statistics*, **32**, 315–326.

7. Self, S. G. and Grossman, E. A. (1986). Linear rank tests for interval–censored data with application to PCB levels in adipose tissue of transformer repair workers, *Biometrics*, **42**, 521–530.

8. Williams, J. S. and Lagakos, S. W. (1977). Models for censored survival analysis: constant–sum and variable–sum models, *Biometrika*, **64**, 215–224.

16

Multivariate Survival Data With Censoring

Shulamith Gross and Catherine Huber-Carol

Baruch College of the City University of New York, Dept of Statistics and CIS, Box 11-220, 1 Baruch way, 10010 NY.
Université Paris V, René Descartes, 45 rue des Saints-Pères, 75 006 Paris, and INSERM U 780.

Abstract: We define a new class of models for multivariate survival data, in continuous time, based on a number of cumulative hazard functions, along the lines of our family of models for correlated survival data in discrete time (Gross and Huber, 2000, 2002). This family is an alternative to frailty and copula models. We establish some properties of our family and compare it to Clayton's and Marshall-Olkin's. Finally we derive non parametric partial likelihood estimates of the hazards involved in its definition and prove, using martingale theory, their asymptotic normality. Simulations will be performed as well as applications to diabetic retinopathy and tumorigenesis in rats.

Keywords and phrases: Survival data, clusters, right censoring, continuous time, hazard rates

16.1 Introduction

Much attention has been paid to multivariate survival models and inference since the early work of Hougaard, and his recent book (2004) on the subject. Studies on twins lead to the development of papers on bivariate distributions, and, more generally the analysis of family data or clusters data lead to more general models for correlated survival data. One way of dealing with this problem is to use copula or frailty models (see for example Bagdonavicius and Nikulin (2002) for a review of those models). Among the most usual bivariate models, one finds Clayton's, Marshall-Olkin's and Gumbel's models. We shall present here a model for continuous multivariate data based on the same idea as the one we used in the discrete case (Gross and Huber, (2002)), and which is closely related to a multi-state process. We define our class of models in detail for the special case of bivariate data, and generalize this class to any dimension. We then obtain properties of these models and compare them to the usual ones

cited above. We then derive NPML estimators for the involved functions and derive their asymptotic properties.

16.2 Definition of the models

16.2.1 Bivariate continuous model

Let $\mathcal{L}$ be the class of continuous univariate cumulative hazard functions on $\mathbb{R}^+$:

$$\mathcal{L} = \{\Lambda : \mathbb{R}^+ \to \mathbb{R}^+, \text{ continuous, non decreasing }, \Lambda(0) = 0, \Lambda(t) \xrightarrow[t\to\infty]{} \infty\}$$

Definition 1 (bivariate continuous model)
Given any five members $\Lambda_{11}^{01}, \Lambda_{11}^{10}, \Lambda_{11}^{00}, \Lambda_{01}^{00}, \Lambda_{10}^{00}$ of $\mathcal{L}$, we define a joint bivariate survival function S on $\mathbb{R}^+ \times \mathbb{R}^+$ by

$$\begin{aligned}
\text{for } x < y,\ dS(x,y) &= exp\{-\Lambda_{11}^{01}(x) - \Lambda_{11}^{10}(x) - \Lambda_{11}^{00}(x)\} d\Lambda_{11}^{01}(x) \\
&\quad exp\{-(\Lambda_{01}^{00}(y) - \Lambda_{01}^{00}(x)\} d\Lambda_{01}^{00}(y) \\
\text{for } y < x,\ dS(x,y) &= exp\{-\Lambda_{11}^{01}(y) - \Lambda_{11}^{10}(y) - \Lambda_{11}^{00}(y)\} d\Lambda_{11}^{10}(y) \quad (16.1) \\
&\quad exp\{-(\Lambda_{10}^{00}(x) - \Lambda_{10}^{00}(y)\} d\Lambda_{10}^{00}(x) \\
\text{for } y = x,\ dS(x,y) &= exp\{-\Lambda_{11}^{01}(x) - \Lambda_{11}^{10}(x) - \Lambda_{11}^{00}(x)\} d\Lambda_{11}^{00}(x)
\end{aligned}$$

We propose the family (16.1) of bivariate probabilities as an alternative to the bivariate probabilities defined by frailties or copulas. It is easy to verify that S thus defined is actually a bivariate survival function, and that a necessary and sufficient condition for the corresponding probability to be absolutely continuous (AC) with respect to λ^2, the Lebesgue measure on $\mathbb{R}^2$, is that $\Lambda_{11}^{00} \equiv 0$. Otherwise, part of the mass is on the diagonal of $\mathbb{R}^2$.

16.2.2 Generalization to p components

When more than two components are involved, say p, then our class of models is defined in a similar way, involving now a number of cumulative hazards $K(p)$ equal to

$$K(p) = \sum_{k=0}^{p-1} C_p^{p-k} C_{p-k}^1. \tag{16.2}$$

when the multivariate law is absolutely continuous with respect to λ^p, the Lebesgue measure on $\mathbb{R}^p$, and

$$K(p) = \sum_{k=0}^{p-1} C_p^{p-k} (2^{p-k} - 1). \tag{16.3}$$

when simultaneous jumps are allowed.

16.2.3 Properties of the bivariate family

Theorem 1 *For all bivariate survival functions defined above and such that* $\Lambda_{11}^{00} \equiv 0$*, we have the following conditional hazard rates* $\forall s < t \in I\!R^+$*:*

$$\begin{array}{lcll} P(X = dt, Y > t | X \geq t, Y \geq t) & = & d\Lambda_{11}^{01}(t) & \\ P(X > t, Y = dt | X \geq t, Y \geq t) & = & d\Lambda_{11}^{10}(t) & \\ P(X = dt | X \geq t, Y < t) & = & d\Lambda_{10}^{00}(t) & = \quad P(X = dt | X \geq t, Y = ds) \\ P(Y = dt | Y \geq t, X < t) & = & d\Lambda_{01}^{00}(t) & = \quad P(Y = dt | Y \geq t, X = ds) \end{array}$$

Conversely, if there exist $\Lambda_{11}^{10}, \Lambda_{11}^{01}, \Lambda_{10}^{00}, \Lambda_{01}^{00}$*, cumulative hazard functions in* $\mathcal{L}$ *such that the joint law satisfies the above equations, then the joint survival function of* (X, Y) *satisfies (16.1).*

Theorem 2 *If* (X, Y) *has survival function* S *given by (16.1), then* X *and* Y *are independent and* S *is absolutely continuous with respect to* λ^2 *if and only if*

$$\Lambda_{11}^{00} \equiv 0 \ ; \ \Lambda_{11}^{01} \equiv \Lambda_{10}^{00} \ ; \ \Lambda_{11}^{10} \equiv \Lambda_{01}^{00}.$$

A version of our model (16.1), in discrete time, was introduced in Gross and Huber (2000). The two models are embedded in the general model obtained by replacing, in (16.1), $\mathcal{L}$ by $\mathcal{L}^*$, the set of cumulative hazards with possible jumps on an at most denumerable set of points $\mathcal{D} \in I\!R^+$:

$$\mathcal{L}^* = \{\Lambda : I\!R^+ \to I\!R^+, \Lambda \text{ non decreasing }, \Lambda(0) = 0, \Lambda(t) \xrightarrow[t \to \infty]{} \infty\}$$

Simple examples of laws of type (16.1) may be obtained by choosing usual parametric hazards for the involved $\Lambda' s$.

16.3 NPML estimation

16.3.1 Likelihood for the bivariate case

Let $X = (X_{i1}, X_{i2})$ be the bivariate survival time of cluster i, $i \in \{1, 2, \ldots, n\}$. The clusters are assumed to be independent. X_{i1} and X_{i2} may possibly be right censored by a bivariate censoring time $C = (C_{i1}, C_{i2})$, independent of X, so that the observed bivariate time is $T = ((X_{i1} \wedge C_{i1}, X_{i2} \wedge C_{i2}) \equiv (T_{i1}, T_{i2})$. The indicator of non censoring is denoted $\delta = (\delta_{i1}, \delta_{i2}) \equiv (1\{T_{i1} = X_{i1}\}, 1\{T_{i2} = X_{i2}\})$. Let $R_{ij}(t) = 1\{t < T_{ij}\}$ and $N_{ij}(t) = \delta_{ij} 1\{t \geq T_{ij}\}$ be respectively the associated at risk and counting processes defined for $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2\}$, and $R(t) = (R_{i1}(t), R_{i2}(t))$, $N(t) = (N_{i1}(t), N_{i2}(t))$. The likelihood will be expressed in terms of the hazards defined as $\lambda_{11}^{01}(t)dt = P(t \leq X_1 \leq$

$t + dt | X_1 \geq t, X_2 > t)$ and the like. It is the product $V = \prod_{i=1}^{n} V_i$ where each V_i may be written as

$$\begin{array}{rcl} V_i & = & \prod_t (1 - \lambda_{11}^{10}(t)dt - \lambda_{11}^{01}(t)dt)^{R_1(t)R_2(t)} (\lambda_{11}^{10}(t))^{R_1(t^-)R_2(t^-)dN_1(t)} \\ & & (\lambda_{11}^{01}(t))^{R_1(t^-)R_2(t^-)dN_2(t)} \prod_t (1 - \lambda_{10}^{10}(t)dt)^{R_1(t)(1-R_2(t))\delta_2} \\ & & \prod_t (1 - \lambda_{01}^{01}(t)dt)^{R_2(t)(1-R_1(t))\delta_1} (\lambda_{10}^{10}(t)dt)^{R_1(t)(1-R_2(t))\delta_2 dN_1(t)} \\ & & (\lambda_{01}^{01}(t)dt)^{R_2(t)(1-R_1(t))\delta_1 dN_2(t)} \end{array}$$

Maximization of V (NPML) implies jumps of the $\Lambda' s$ at (ordered) times T_k , $k = 1, 2 \cdots, K$ when an event occurred ($\delta_{ij} = 1$ for some (i, j)). Let us introduce the quantities:

$$\begin{array}{rclcrcl} \tau_1(i) & = & 1\{T_{i1} < T_{i2}\} \; ; \; \tau_2(i) = 1\{T_{i2} < T_{i1}\} & ; & \tau(i) & = & 1\{T_{i1} = T_{i2}\} \\ a_k & = & \Lambda_{11}^{01}(T_k^+) - \Lambda_{11}^{01}(T_k^-) & ; & b_k & = & \Lambda_{11}^{10}(T_k^+) - \Lambda_{11}^{10}(T_k^-) \; ; \\ c_k & = & \Lambda_{10}^{00}(T_k^+) - \Lambda_{10}^{00}(T_k^-) & ; & d_k & = & \Lambda_{01}^{00}(T_k^+) - \Lambda_{01}^{00}(T_k^-) \; . \end{array}$$

and the counts:

$$\begin{array}{rcl} s_1(i) = \sum_{i'} 1\{T_{i1} \leq T_{i'1} \wedge T_{i'2}\} & ; s_2(i) = \sum_{i'} 1\{T_{i2} \leq T_{i'1} \wedge T_{i'2}\} & ; \\ s_3(i) = \sum_{i'} \tau_2(i') 1\{T_{i'2} \leq T_{i1} \leq T_{i'1}\}\} & ; s_4(i) = \sum_{i'} \tau_1[i'] 1\{T_{i'1} \leq T_{i2} \leq T_{i'2}\}\}. \end{array}$$

Then the log-likelihod is equal to

$$\begin{array}{rcccccl} L & = & -\sum_i a_i \delta_{i1} \tau_1(i) s_1(i) & - & \sum_i b_i \delta_{i2} \tau_2(i) s_2(i) & + & \sum_i \delta_{i1} \tau_1(i) \log(a_i) \\ & & \sum_i \delta_{i2} \tau_2(i) \log(b_i) & - & \sum_i c_i \delta_{i1} \tau_2(i) b_i s_3(i) & - & \sum_i d_i \delta_{i2} \tau_1(i) b_i s_4(i) \\ & & \sum_i \delta_{i1} \tau_2(i) \log(c_i) & + & \sum_i \delta_{i2} \tau_1(i) \log(d_i) & & \end{array}$$

By derivation of L with respect to a_i, b_i, c_i, d_i, we obtain the NPML estimates:

$$\widehat{a_i} = \frac{\delta_{i1}\tau_1(i)}{s_1(i)} \; ; \; \widehat{b_i} = \frac{\delta_{i2}\tau_2(i)}{s_2(i)} \; ; \; \widehat{c_i} = \frac{\delta_{i1}\tau_2(i)}{s_3(i)} \; ; \; \widehat{d_i} = \frac{\delta_{i2}\tau_1(i)}{s_4(i)} \; .$$

In order to derive their asymptotic properties, one rewrites them in terms of the associated counting N, at risk Y and martingale M processes with respect to the filtration $\mathcal{F}(t) = \sigma(N_{i1}(s), N_{i2}(s), R_{i1}(s), R_{i2}(s), s < t)$, for each case: jump of individual 1 (resp. 2) in the presence (resp. absence) of the other element of the pair:

$$\begin{array}{rcccl} N_{i,11:01}(t) & = & 1\{X_{i1} \leq t, X_{i1} < X_{i2} \wedge C_{i1} \wedge C_{i2}\} & = & \int_0^t R_{i1}(s) R_{i2}(s) dN_{i1}(s) \\ Y_{i,11}(t) & = & 1\{X_{i1} \wedge X_{i2} \wedge C_{i1} \wedge C_{i2} \geq t\} & = & R_{i1}(t) R_{i2}(t) \\ M_{i,11:01}(t) & = & N_{i,11:01}(t) - \int_0^t Y_{i,11}(u) d\Lambda_{11}^{01}(u) & & \end{array}$$

and the like, $(N_{i,11:10}(t), Y_{i,11}(t), M_{i,11:10}(t))$,$(N_{i,10:00}(t), Y_{i,10}(t), M_{i,10:00}(t))$, and $(N_{i,01:00}(t), Y_{i,01}(t), M_{i,01:00}(t))$. The whole asymptotic normal theory holds as the estimates of the cumulative $\Lambda' s$, properly normalized converge to independent gaussian martingales with estimable covariances.

16.4 Concluding remarks

The proposed model could be considered as a multistate model, where the successive states are the actual composition of the subset of the cluster that is *still at risk* after some members have experienced the expected event. In a future work, we shall introduce covariates such as cluster and individual covariates as well as the time elapsed between two successive states of the cluster. Let us finally remark that the parallel with semi-Markov models for multistate models is not straightforward. This is due to the fact that, for example in the bivariate case, when the pair is in state $(0, 1)$ the cumulative hazard Λ_{01}^{00} starts from 0 and not from the time s at which the first member of the pair experienced the event. Making the parallel perfect would lead to a new family of models having all properties of semi-markov multistate models, to which could be applied all results already obtained for example by Huber, Pons and Heutte (2006).

References

1. Bagdonavicius, V. and Nikulin, M. (2002). *Accelerated Life Models*, Chapman and Hall/CRC, Boca Raton.

2. Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65** , 543–52.

3. Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion), *Journal of the Royal Statistical Society A*, **148** , 82–117.

4. Genest, C., Ghoudi, K. and Rivest L-P (1995). A semi-parametric estimation procedure of dependence parameter in multivariate families of distributions, *Biometrika*, **82(3)** , 543–552.

5. Gross, S. and Huber-Carol, C. (2000). Hierarchical Dependency Models for Multivariate Survival data with Censoring, *Lifetime Data Analysis***6**, 299–320.

6. Gross, S. and Huber-Carol, C. (2002). A new family of Multivariate Distributions for Survival Data, In *Goodness of Fit tests and Model Validity*, (Ed. C. Huber-Carol, N. Balakrishnan, M.S. Nikulin, M. Mesbah), 255–266, Birkhauser.

7. Hougaard, P. (2004). *Analysis of Multivariate Survival Data*, Springer Verlag ed.

8. Huber-Carol, C., Pons, O. and Heutte N. (2006). Inference for a general semi-Markov model and a sub-model for independent competing risks, In *Probability, Statistics and Modelling in Public Health* (Ed., M. Nikulin, D. Commenges, C. Huber), pp. 231–244, Springer.

9. Lee, Wei, and Ying (1993) Linear regression analysis for highly stratified failure time data, *Biometrics*, **58** , 643–649.

10. Marschall, A.W. and Olkin, I. (1967). A multivariate exponential distribution, *J.A.S.A.*,**62**, 30–44.

11. Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association*, **84** , 487–493.

12. Ross, E.A. and Moore, D. (1999) Modeling clustered, discrete, or grouped time survival data with covariates, *Biometrics*, **55(3)** , 813–819.

17

Goodness of Fit Tests for Pareto Distribution

Sneh Gulati and Samuel Shapiro

Department of Statistics
Florida International University

Abstract: The Pareto Distribution can serve to model several types of data sets, especially those arising in the insurance industry. In this paper, we present methods to test the hypothesis that the underlying data come from a Pareto Distribution. The test presented for the Type I Pareto Distribution is based on the regression test of Brain and Shapiro (1983) for the exponential distribution while the test for the Type II Pareto Distribution uses the modified Greenwood statistic developed by Shapiro and Chen (2001). Power comparisons of the tests are carried out via simulations.

Keywords and phrases: Type I Pareto distribution, Type II Pareto distribution, regression tests, Greenwood's statistic, extreme

17.1 Introduction

Statisticians and engineers have been expanding the types of models used in the analysis of measurement data. There was a time when the normal distribution was used as an underlying model for all data, but of late the exponential, Weibull, lognormal, gamma and Pareto distributions have been used in the search for models that more closely match the phenomena under study. Since the choice of a model can significantly affect the results of the analysis of a data set, testing model assumptions plays an important role in such analysis. This paper presents composite tests for the assumption that a set of data comes from a Pareto population.

The Pareto distribution originates from the work of Pareto (1897) and has been used in many applications including modeling income distributions, hy-

drology, insurance claims and in general, populations representing extreme occurrences. Arnold (1983) stated that this model was useful for approximating data that arose from distributions with "fat tails". A comprehensive discussion of the Pareto distribution can be found in this reference. Various modifications have been made to the classical distribution proposed by Pareto. In Arnold's book he has labelled these as Type I, Type II, Type III and Type IV. This article will discuss a distributional test for the first two of these distributions. The estimation of the Pareto parameters using the maximum likelihood method results in biased estimates for the Type I Pareto and is not straightforward for the Type II. Several authors have presented simplified corrections for the bias of the maximum likelihood estimators for the Type I model (see for e.g., Saksena (1978), Baxter (1980), and Cook and Mumme (1981)). Estimation for the parameters of the Type II model has been studied extensively as well. Harris (1968) and Arnold and Laguna (1977) proposed the technique of matching of moments for the Type II while Davison (1984) and Grimshaw (1993) (among others) have studied maximum likelihood estimation for the Type II Pareto.

The classical or type I Pareto distribution is defined by the density function,

$$f(y) = \frac{\alpha\sigma^{\alpha}}{y^{\alpha+1}}, \quad y \geq \sigma > 0,\ \alpha > 0. \tag{1.1}$$

The parameter α is the shape parameter and σ is the scale parameter. Note that the minimum value of Y is equal to σ. It is easy to see that if Y has a Type I Pareto distribution, then $T = ln(Y/\sigma)$ has an exponential distribution with a mean of $1/\alpha$.

The Type II Pareto distribution is defined by the distribution function

$$F(y) = 1 - \left(1 + \frac{y}{\sigma}\right)^{-\alpha}, \quad y \geq 0,\ \sigma \geq 0,\ \alpha > 0. \tag{1.2}$$

If one assumes that $\alpha > 1$ then the distribution has a finite first moment.

Unlike other distributions there are few tests to assess whether it is reasonable to use the Pareto model with a given set of data when the two parameters are unknown. When the parameters are known it is a simple matter to transform the data to an exponential distribution and use one of the many tests for the exponential. In the composite case one can use the classical chi squared goodness of fit procedure using maximum likelihood estimates of the parameters; but this procedure usually has poor power properties for continuous distributions. Choulakian and Stephens (2001) developed two composite hypothesis tests for a generalized Pareto distribution based on the Anderson-Darling and the Cramer-von Mises statistics. However, the generalized Pareto is a Type II distribution that has been parametrized and where the parameter α can be negative. We were interested in investigating testing procedures for the classical Type I and Type II Pareto models. For the Type I Pareto model, this paper proposes a composite hypothesis test that is based on transforming the

data to an exponential distribution and uses a modification of a test of exponentiality devised by Brain and Shapiro (1983). The test statistic proposed by Brain and Shapiro has an asymptotic chi-squared distribution with two degrees of freedom distribution and requires a correction factor that is a function of the sample size. Simulation studies have indicated that the null distribution of the proposed test statistic for the Pareto I distribution can be approximated by the chi squared distribution for sample sizes as small as 10 without the correction. For the Type II Pareto, we propose a test statistic based on the procedure developed by Shapiro and Chen (2001) for testing the gamma distribution.

We now describe the testing procedures.

17.2 Type I Pareto Distribution

The first step in testing for either the Type I or the Type II Pareto distribution is the estimation of the parameters. Maximum likelihood estimates for the Type I Pareto are easy to compute and are given by:

$$\sigma_{mle} = Y_{(1)} = \text{the smallest observation} \tag{2.1}$$

$$\alpha_{mle} = \frac{n}{\sum_{i=1}^{n} \ln\left(\frac{Y_{(i)}}{Y_{(1)}}\right)}. \tag{2.2}$$

The estimator of α is biased which can be corrected by multiplying by $(n-2)/n$. Both these estimators are consistent and mutually independent and their sampling distributions are given in Malik (1970). A simulation study of the mean square error for the shape and scale parameters shows that the MSE's are usually quite small (below 0.01 in most cases) indicating that using the scale parameter estimate to transform the data to an exponential distribution should yield satisfactory results.

Based on the regression test of Brain and Shapiro (1983) for the exponential distribution, the test for the Type I Pareto distribution is then conducted as follows:

1. Denote the data values by $Y_1, Y_2, ..., Y_n$
2. Obtain the estimate of the scale parameter using the maximum likelihood estimator from (2.1)
3. Transform the data to $W_i = \ln(X_i)$ where $X_i = \frac{Y_i}{\sigma_{mle}} = \frac{Y_i}{Y_{(1)}}$. Note that this transformation converts the shape parameter to a scale parameter and the scale parameter to the origin parameter of zero. As pointed out in Section 1, the transformed data will be exponential with origin of zero and a scale parameter of α.
4. Order the transformed variables and compute the $(n-1)$ weighted spacings; $V_i = (n-i+1)(W_{(i)} - W_{(i-1)}),\ i = 1, 2, \ldots, n$ and $V_0 = 0$. Note that $W_{(1)} \leq W_{(2)} \leq ... \leq W_{(n)}$ are the order statistics of the W_i's.

5. Compute

$$Z_1 = \sqrt{\frac{12}{n-1}}\Big(\frac{\sum_{i=1}^{n-1}\alpha_i V_{i+1}}{\sum_{i=1}^{n-1} V_{i+1}}\Big) \tag{2.3}$$

where $a_i = i - n/2,\ i = 1, 2, \ldots, n-1$.
6. Compute

$$Z_2 = \sqrt{\frac{5}{4(n+1)(n-2)(n-3)}}\frac{12\sum_{i=1}^{n-1}\alpha_i^2 V_{i+1} - n(n-2)\sum_{i=1}^{n-1} V_{i+1}}{\sum_{i=1}^{n-1} V_{i+1}} \tag{2.4}$$

7. Finally the test statistic

$$Z = Z_1^2 + Z_2^2 \tag{2.5}$$

8. The limiting distribution of Z is a chi square distribution with 2 degrees of freedom, so reject the null hypothesis of a Pareto Type I distribution if $Z > \chi^2_{2,\alpha}$ for an α level test.

The power of the test has been tested against various alternatives and the simulation results show that the test is extremely powerful with the power being higher than 0.70 in most cases.

17.3 Type II Pareto Distribution (Work in Progress)

Once again the first step in testing for the Type II Pareto Distribution involves the estimation of the underlying parameters, σ and α. The maximum likelihood estimates of the parameters are obtained by solving the following non linear equation for σ :

$$f(\hat{\sigma}) = N^2 - \big(\sum_{i=1}^{N}\frac{\hat{\sigma}}{\hat{\sigma}+X_i}\big) \star \big(N - \sum_{i=1}^{N}\ln\frac{\hat{\sigma}}{\hat{\sigma}+X_i}\big) = 0. \tag{3.1}$$

Once we have $\hat{\sigma}$, the estimate of α, $\hat{\alpha}$ is given by:

$$\hat{\alpha} = \frac{N}{\sum_{i=1}^{N}\ln(\hat{\sigma}+X_i) - n\ln\hat{\sigma}}. \tag{3.2}$$

To solve equation (3.1), we will use the iterative method proposed by Grimshaw (1993). Provided that the estimators have a small mean square error, the modified Greenwood statistic as developed by Shapiro and Chen (2001) will be used to test for the Type II Pareto distribution. The procedure is described as follows:
1. Denote the data values by $Y_1, Y_2, ..., Y_n$
2. Obtain the estimates of the scale parameter and the shape parameters, $\hat{\sigma}$ and α respectively, using (3.1) and (3.2)
3. Transform the data to $Z_i = 1 - [1 + Y_i/\hat{\sigma}]^{-\hat{\alpha}}$. Note that the transformed data will be made up of i.i.d. uniform (0,1) random variables.

4. Order the transformed data and compute the (n+1) spacings; $D_i = Z_{(i)} - Z_{(i-1)}, i = 1, 2, ..., n+1$ where $Z_{(1)} \leq Z_{(2)} \leq ... \leq Z_{(n)}$ are the ordered Z values and $Z_{(0)} = 0, Z_{(n+1)} = 1$.
5. Compute the modified Greenwood statistic, G_w as follows:

$$G_w = (n+1) \sum_{i=1}^{n+1} \alpha_i D_i^2 \tag{3.3}$$

where $a_1 = a_2 = a_3 = a_4 = 3/2(n+1)$ and $a_i = \frac{1-4\alpha_1}{n-3},\ i = 3, 4, ..., n-1$.
6. The percentiles and the rejection region for the test will be determined via Monte Carlo simulations since the asymptotic distribution of G_w is not available.

Finally the power of the proposed test will also be studied via simulations.

References

1. Arnold, B. C. (1983). *Pareto Distributions*, Statistical Distributions in Scientific Work Series, Vol. 5.

2. Arnold, B. C. and Laguna, L.(1977). *On Generalized Pareto Distributions with Applications to Income Data*, International Studies in Economics, Monograph #10, Department of Economics, Iowa State University, Iowa.

3. Baxter, M. A., (1980). Minimum Variance Unbiased Estimators of the Parameters of the Pareto Distribution, *Metrika*, 27,133-138.

4. Brain, C. W. and Shapiro, S. S. (1983). A Regression Test for Exponentiality: Censored and Complete Samples, *Technometrics*, 25, 69-76.

5. Choulakian, V. and Stephens, M. A. (2001). Goodness-of-Fit Tests for the generalized Pareto Distribution, *Technometrics*, 43, 478-484.

6. Cook, W. L, and Mumme, D.C. (1981). Estimation of Pareto Parameters by Numerical Methods, *Statistical Distributions in Scientific Work*, Vol. 5, Taillie C., Patil, G.P. and Baldessari, B., eds., Reidel, Dordecht-Holland, 127-132.

7. Cox, D. R. and Lewis, P. A. (1966). *The Statistical Analysis of a Series of Events*, London: Methune.

8. Davison A.C. (1984). Modeling Excesses over High Thresholds, With an Application, in *Statistical Extremes and Applications*, ed. J. Tiago de Oliveira, Dordrecht: D Reidel, 461-482.

9. Grimshaw S.D. (1993). Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution, *Technometrics*, 35, 185-191.

10. Harris, C. M. (1968). The Pareto Distribution as a Queue Service Discipline, *Operations Research*, 16, 307-313.

11. Malik, H. J. (1970). Estimation of the Parameters of the Pareto Distribution, *Metrika*, 15, 126-132.

12. Pareto, V. (1897). *Cours d'economie Politique*, Vol. II. F. Rouge, Lausanne.

13. Saksena, S. K. (1978). *Estimation of Parameters in a Pareto Distribution and Simultaneous Comparison of Estimators*, Ph.D. Thesis, Louisiana Tech University, Ruston, Louisiana.

14. Shapiro, S.S. and Chen, L. (2001). Composite Tests for the Gamma Distribution, *Journal of Quality Technology*, 33(1), 47-59.

18

Focussed Information Criteria for the Linear Hazard Regression Model

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract: The linear hazard regression model developed by Aalen is becoming an increasingly popular alternative to the Cox multiplicative hazard regression model. There are no methods in the literature for selecting among different candidate models of this nonparametric type, however. In the present paper a focussed information criterion is developed for this task. The criterion works for each specified covariate vector, by estimating the mean squared error for each candidate model's estimate of the associated cumulative hazard rate; the finally selected model is the one with lowest estimated mean squared error. Averaged versions of the criterion are also developed.

Keywords and phrases: Aalen's linear model, covariate selection, focussed information criterion, hazard regression, model selection

18.1 Introduction: Which Covariates to Include?

We consider survival regression data of the usual form (T_i, δ_i, x_i) for individuals $i = 1, \ldots, n$, where x_i is a vector of say r covariates, among which one wishes to select those of highest relevance. Also, $T_i = \min\{T_i^0, C_i\}$ is the possibly censored life-length and $\delta_i = I\{T_i^0 < C_i\}$ the associated non-censoring indicator, in terms of underlying life-length T_i^0 and censoring time C_i for individual i.

Our framework is that of the linear hazard regression model introduced by Aalen (1980), see e.g. the extensive discussion in Andersen, Borgan, Gill and Keiding (1993, Ch. 8), where the hazard rate for individual i may be represented as

$$h_i(u) = x_i^{\mathrm{t}}\alpha(u) = \sum_{j=1}^{r} x_{i,j}\alpha_j(u) \quad \text{for } i = 1, \ldots, n,$$

in terms of regressor functions $\alpha_1(u), \ldots, \alpha_r(u)$. These need to satisfy the requirement that the linear combination $x^{\mathrm{t}}\alpha(u)$ stays nonnegative for all x supported by the distribution of covariate vectors. In other words, the associated cumulative hazard function

$$H(t \,|\, x) = \int_0^t x^{\mathrm{t}}\alpha(u)\,\mathrm{d}u = x^{\mathrm{t}}A(u) = \sum_{j=1}^r x_j A_j(t) \tag{18.1}$$

is nondecreasing in t, for all x in the relevant covariate space; here we write $A_j(t) = \int_0^t \alpha_j(u)\,\mathrm{d}u$ for $j = 1, \ldots, r$.

Among questions discussed in this paper is when we might do better with only a subset of the x covariates than with keeping them all. We focus specifically on the problem of estimating $H(t \,|\, x)$ of (18.1) well, for a specified individual carrying his given covariate information x. The full-model estimator

$$\widehat{H}(t \,|\, x) = \widehat{H}_{\mathrm{full}}(t \,|\, x) = x^{\mathrm{t}}\widehat{A}(t) = \sum_{j=1}^r x_j \widehat{A}_j(t) \tag{18.2}$$

is one option, using the familiar nonparametric Aalen estimators for $A_1, \ldots, A_r$ in the full model, keeping all covariates on board. Pushing some covariates out of the model leads to competing estimators of the type

$$\widetilde{H}_I(t \,|\, x) = \sum_{j \in I} x_j \widetilde{A}_{I,j}(t), \tag{18.3}$$

where the index set I is a subset of $\{1, \ldots, r\}$, representing those covariates that are kept in the model, and where the $\widetilde{A}_{I,j}(t)$'s for $j \in I$ are the Aalen estimators in the linear hazard rate model associated with the I covariates. Using $\widetilde{H}_I(t \,|\, x)$ instead of $\widehat{H}(t \,|\, x)$ will typically correspond to smaller variances but to modelling bias. Slightly more generally, bigger index sets I imply more variance but less modelling bias, and vice versa. Thus the task of selecting suitable covariates amounts to a statistical balancing game between sampling variability and bias.

In the present 'extended abstract' we can only briefly report on developments, findings and applications presented more fully in the technical report Hjort (2006). In Section 18.2 we fix the framework and give proper definitions of full-model and submodel estimators, partly in terms of counting processes and at-risk processes. Links with martingale theory make it possible to reach results reported on in Section 18.3 that accurately describe the bias and variance properties associated with a given candidate model. These quantities can then be estimated from data. The focussed information criterion (FIC) introduced in Section 18.4 acts by estimating the risk associated with each candidate model's estimator of the cumulative hazard function; the model we suggest being used

in the end is the one with lowest estimated risk. A weighted version is also put forward.

This brief introduction has so far taken model comparison as corresponding to accuracy of estimators of cumulative hazard rates $H(t \,|\, x)$. By a delta method argument this is also nearly equivalent to ranking models in terms of accuracy of estimates of survival probabilities $S(t \,|\, x) = \exp\{-H(t \,|\, x)\}$. It is important to realise that a submodel I may work better than the full model, even if the submodel in question is not 'fully correct' as such; this is determined, among other aspects, by the sizes of the $\alpha_j(u)$ regressor functions that are left out a model. This makes model selection different in spirit and operation than e.g. performing goodness-of-fit checks on all candidate models.

General methodology for focussed information criteria and frequentist model average inference has been developed in Hjort and Claeskens (2003) and Claeskens and Hjort (2003) for the case of likelihood theory for parametric models. The work reported on here may be seen as suitable extensions of that methodology, to the present framework of nonparametric hazard regression models; see again Hjort (2006) for a fuller account.

18.2 Estimators in Submodels

This section properly defines the Aalen estimators $\widehat{A}$ and $\widetilde{A}_I$ involved in (18.2) and (18.3). It is convenient to define these in terms of the counting process and at-risk process

$$N_i(t) = I\{T_i \le t, \delta_i = 1\} \quad \text{and} \quad Y_i(u) = I\{T_i \ge u\}$$

for individuals $i = 1, \ldots, n$. Now introduce the $r \times r$-size matrix function $G_n(u) = n^{-1} \sum_{i=1}^n Y_i(u) x_i x_i^{\mathrm{t}}$. The Aalen estimator $\widehat{A} = (\widehat{A}_1, \ldots, \widehat{A}_r)^{\mathrm{t}}$ is

$$\widehat{A}(t) = \int_0^t G_n(u)^{-1} n^{-1} \sum_{i=1}^n x_i \, \mathrm{d}N_i(u) \quad \text{for } t \ge 0.$$

This also defines $\widehat{H}_{\mathrm{full}}(t \,|\, x)$ of (18.2). It is assumed here that at least r linearly independent covariate vectors x_i remain in the risk set at time t, making the inverse of G_n well-defined for all $u \le t$; this event has probability growing exponentially quickly to 1 as sample size increases, under mild conditions.

To properly define the competitor $\widetilde{H}_I(t \,|\, x)$ of (18.3), we use the notation $x_I = \pi_I x$ for the vector of those x_j components for which $j \in I$, for each given subset I of $\{1, \ldots, r\}$. In other words, π_I is the projection matrix of size $|I| \times r$, with $|I|$ the number of covariates included in I. For the given I, we partition the G_n function into blocks,

$$G_n(u) = \begin{pmatrix} G_{n,00}(u), & G_{n,01}(u) \\ G_{n,10}(u), & G_{n,11}(u) \end{pmatrix}.$$

where $G_{n,00}(u) = n^{-1}\sum_{i=1}^n Y_i(u)x_{i,I}x_{i,I}^{\rm t}$ is of size $|I|\times|I|$, etc. The Aalen estimator for the vector of A_j functions where $j\in I$ is

$$\widetilde{A}_I(t) = \int_0^t G_{n,00}(u)^{-1}n^{-1}\sum_{i=1}^n x_{i,I}\,{\rm d}N_i(u).$$

These are those at work in (18.3).

Using martingale theory for counting processes one may express ${\rm d}\widetilde{A}_I(u)$ as ${\rm d}A_I(u)+G_{n,00}(u)^{-1}G_{n,01}(u)\,{\rm d}A_{II}(u)$ plus martingale noise. In particular, when the I model is used, then the Aalen estimator $\widetilde{A}_I(t)$ does not really estimate $A_I(t)$, but rather the function $A_I(t)+\int_0^t G_{00}^{-1}G_{01}\,{\rm d}A_{II}$, where G_{00} and so on are limit versions of $G_{n,00}$ and so on; see below.

18.3 Assessing and Estimating Bias, Variance, and Mean Squared Error

In this section we first indicate how useful approximations can be developed for the mean squared error of each of the (18.3) estimators $\widetilde{H}_I(t\,|\,x) = x_I^{\rm t}\widetilde{A}_I(t)$, and then use these to construct natural risk estimators. We shall assume that the censoring variables $C_1,\ldots,C_n$ are i.i.d. with some survival distribution $C(u) = \Pr\{C_i\ge u\}$, and that they are independent of the life-times T_i^0; the case of no censoring corresponds to $C(u)=1$ for all u. It will furthermore be convenient to postulate that $x_1,\ldots,x_n$ stem from some distribution in the space of covariate vectors. These assumptions imply for example that the G_n function converges with increasing sample size, say

$$G_n(u)\to G(u) = {\rm E}_* Y(u)xx^{\rm t} = {\rm E}_*\exp\{-x^{\rm t}A(u)\}xx^{\rm t}\,C(u), \tag{18.4}$$

where ${\rm E}_*$ refers to expectation under the postulated covariate distribution. Also the mean function $\bar{G}_n(u) = n^{-1}\sum_{i=1}^n p_i(u)x_ix_i^{\rm t}$ converges to the same limit $G(u)$; here $p_i(u) = {\rm E}Y_i(u) = \exp\{-x_i^{\rm t}A(u)\}\,C(u)$. We shall finally assume that the $r\times r$-function $G(u)$ is invertible over the time observation window $u\in[0,\tau]$ of interest; this corresponds to $C(\tau)$ positive and to a non-degenerate covariate distribution. As in Section 18.2 there will be a need to partition the $G(u)$ function into blocks $G_{00}(u), G_{01}(u)$, etc.; $G_{00}(u)$ has e.g. size $|I|\times|I|$.

Consider as in Section 18.1 a given individual with covariate information x, and let

$$b_{I,n}(u) = G_{n,10}(u)G_{n,00}(u)^{-1}x_I - x_{II}, \tag{18.5}$$

which can be seen as a bias function, of dimension $q = r-|I|$. One may now show that

$$\sqrt{n}\{x_I^{\rm t}\widetilde{A}_I(t) - x^{\rm t}A(t)\} = \sqrt{n}\int_0^t b_{I,n}^{\rm t}\,{\rm d}A_{II} + x_I^{\rm t}\int_0^t G_{n,00}^{-1}\,{\rm d}V_{n,I}. \tag{18.6}$$

The second term is a zero-mean martingale, in terms of a certain V_n process, while the first term is a bias term, stemming from using model I that does not include all the components. We may use (18.6) to develop good approximations to say $\text{mse}_n(I,t)$, defined as n times mean squared error of the (18.3) estimator. We treat the covariate vectors $x_1,\ldots,x_n$ as given, i.e. our approximations are expressed directly in terms of these. To give the result we need to define $\mathrm{d}J_n(u) = n^{-1}\sum_{i=1}^n Y_i(u)x_ix_i^{\mathrm{t}}\,x_i^{\mathrm{t}}\,\mathrm{d}A(u)$, the increments of a process which has a well-defined limit

$$\mathrm{d}J(u) = \mathrm{E}_*Y(u)xx^{\mathrm{t}}\,x^{\mathrm{t}}\,\mathrm{d}A(u) = \mathrm{E}_*\exp\{-x^{\mathrm{t}}A(u)\}xx^{\mathrm{t}}\,x^{\mathrm{t}}\,\mathrm{d}A(u)\,C(u)$$

under conditions stated earlier. The basic mse decomposition result is that

$$\text{mse}_n(I,t) \doteq \text{sqb}(I,t) + \text{var}(I,t), \tag{18.7}$$

where the variance term is $x_I^{\mathrm{t}}\int_0^t G_{00}^{-1}\,\mathrm{d}J_{00}\,G_{00}^{-1}\,x_I$ and the squared bias term is

$$\text{sqb}(I,t) = n\Big(\int_0^t \bar b_{I,n}^{\mathrm{t}}\,\mathrm{d}A_{II}\Big)^2,$$

where $\bar b_{I,n}$ is as in (18.5) but with $\bar G_n$ replacing G_n.

We have seen that each candidate model I has an associated risk $\text{mse}_n(I,t)$ of (18.7) when estimating the cumulative hazard function using $\widetilde H_I(t\,|\,x)$. Now we deal with the consequent task of estimating these risk quantities from data. For the variance part we use

$$\widehat{\text{var}}(I,t) = x_I^{\mathrm{t}}\int_0^t G_{n,00}^{-1}(u)\,\mathrm{d}\widehat J_{n,00}(u)\,G_{n,00}(u)^{-1}\,x_I,$$

wherein $\mathrm{d}\widehat J_n(u) = n^{-1}\sum_{i=1}^n Y_i(u)x_ix_i^{\mathrm{t}}\,x_i^{\mathrm{t}}\mathrm{d}\widehat A(u)$, engaging the full-model Aalen estimator. The $|I|\times|I|$ block used for the variance estimation is $\pi_I\,\mathrm{d}\widehat J_n(u)\pi_I^{\mathrm{t}}$. For the squared bias part, considerations given in detail in Hjort (2006) show that

$$\widehat{\text{sqb}}(I,t) = n\Big(\int_0^t b_{I,n}^{\mathrm{t}}\,\mathrm{d}\widehat A_{II}\Big)^2 - \int_0^t b_{I,n}^{\mathrm{t}}\,\mathrm{d}\widehat Q_n\,b_{I,n}$$

is a nearly unbiased estimator, in which

$$\mathrm{d}\widehat Q_n(u) = \pi_{II}\{G_n(u)^{-1}\,\mathrm{d}\widehat J_n(u)\,G_n(u)^{-1}\}\pi_{II}^{\mathrm{t}}.$$

These considerations lead to the risk estimator

$$\widehat R(I,t) = \widehat{\text{mse}}_n(I,t) = \max\{\widehat{\text{sqb}}(I,t),0\} + x_I^{\mathrm{t}}\int_0^t G_{n,00}^{-1}\,\mathrm{d}\widehat J_{n,00}\,G_{n,00}^{-1}\,x_I.$$

18.4 The FIC and the Weighted FIC

Here we show how risk estimation methods developed above lead to natural information criteria for model selection. The first such is a *focussed information criterion* (FIC) that works for a given individual and a given time point at which we wish optimal precision for his survival probability estimate. For the given covariate x and time point t we calculate

$$\text{FIC} = \text{FIC}(I, x, t) = \max\{\widehat{\text{sqb}}(I, x, t), 0\} + \widehat{\text{var}}(I, x, t)$$

for each candidate model I, with these terms computed as in the previous section. We note that $b_{I,n}(u)$ of (18.5) depend on x and that the submatrices $G_{n,00}$ and so on of (18.4) depend on I. In the end one selects the model with smallest FIC score.

Note that FIC is sample-size dependent. In a situation with a given amount of non-zero bias $\int_0^t \bar{b}_I^{\rm t}\,{\rm d}A_{II}$, the $\widehat{\text{sqb}}$ component of FIC will essentially increase with n, whereas the variance component remains essentially constant. This goes to show that the best models will tolerate less and less bias as n increases, and for sufficiently large n only the full model (which has zero modelling bias) will survive FIC scrutiny.

There are various variations on the FIC above, including important versions that correspond to suitable weighted risks; see Hjort (2006) for a full description of such wFIC methods. A special case worth recording is when t is fixed and the weight function w is identical to the covariate distribution. It is unknown, but may be approximated with the empirical distribution of covariates $x_1, \ldots, x_n$. This leads to wFIC(I) as a sum of a weighted variance and a weighted squared bias term, specifically with w-$\widehat{\text{var}}(I)$ equal to

$$n^{-1}\sum_{i=1}^n \widehat{\text{var}}(I, x_i, t) = \text{Tr}\Big\{\Big(\int_0^t G_{n,00}^{-1}\,{\rm d}\widehat{J}_{n,00}\,G_{n,00}^{-1}\Big)\Big(n^{-1}\sum_{i=1}^n x_{i,I}x_{i,I}^{\rm t}\Big)\Big\}$$

and w-$\widehat{\text{sqb}}(I)$ that can be expressed as

$$\sum_{i=1}^n \{x_{i,I}^{\rm t}\widehat{B}_I(t) - x_{i,II}^{\rm t}\widehat{A}_{II}(t)\}^2 - n^{-1}\sum_{i=1}^n \int_0^t b_{I,n}(u, x_i)^{\rm t}\,{\rm d}\widehat{Q}_n(u)\,b_{I,n}(u, x_i),$$

where $\widehat{B}_I(t) = \int_0^t G_{n,00}^{-1}G_{n,01}\,{\rm d}\widehat{A}_{II}(u)$.

References

1. Andersen, P.K., Borgan, Ø., Gill, R. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, Heidelberg.

2. Claeskens, G. and Hjort, N.L. (2003). The focused information criterion [with discussion], *Journal of the American Statistical Association*, **98**, 900–916 and 938–945.

3. Hjort, N.L. (2006). Focussed information criteria for the linear hazard regression model, Statistical Research Report, Department of Mathematics, University of Oslo.

4. Hjort, N.L. and Claeskens, G. (2003). Frequentist average estimators [with discussion], *Journal of the American Statistical Association*, **98**, 879–899 and 938–945.

5. Aalen, O.O. (1980). A model for nonparametric regression analysis of counting processes, in *Mathematical Statistics and Probability Theory* (eds. W. Klonecki, A. Kozek and J. Rosinski). Proceedings of the 6th international conference, Wisla, Poland, 1–25.

19

Semi-parametric Regression Models For Interval-censored Survival Data, With and Without Frailty Effects

Philip Hougaard

Biostatistics, H. Lundbeck A/S, Valby, Denmark

Abstract: Interval-censored survival data occur when the time to an event is assessed by means of blood samples, urine samples, X-ray or other screening methods that cannot tell the exact time of change for the disease, but only that the change has happened since the last examination. This is in contrast to standard thinking that assumes that the change happens at the time of the first positive examination. Even though this screening setup is very common and methods to handle such data non-parametrically in the one-sample case have been suggested more than 25 years ago, it is still not a standard method. However, interval-censored methods are needed in order to consider onset and diagnosis as two different things, like when we consider screening in order to diagnose a disease earlier. The reason for the low use of interval-censored methods is that in the non-parametric case, analysis is technically more complicated than standard survival methods based on exact times. The same applies to proportional hazards models. The talk will cover semi-parametric regression models, both of the proportional hazards type and of the corresponding frailty models, with proportional hazards conditional on the gamma distributed frailty. The statistical theory will not be dealt with in detail. The talk will emphasize the applications, using examples from the literature as well as from my own experience regarding development of microalbuminuria among Type 2 diabetic patients.

Keywords and phrases: Frailty; Interval censoring; Non-parametric estimation

19.1 Introduction

Interval-censored survival data refer to survival data, where the times of events are not known precisely; they are only known to lie in given intervals. This is in contrast to the standard survival data setup, which assumes that all event times are either known precisely, or they happen after the end of observation (that is, right-censored data).

Thus, we only know that the event time is in an interval of the form $(L_i, R_i]$, where the left endpoint is the time last seen without the disease, and the right endpoint is the first time seen with the disease.

The inspection times are supposed to be generated independently of the response process and not informative of the parameters governing the response process. The likelihood then has the following form

$$\prod_i \{S_i(L_i) - S_i(R_i)\}. \tag{19.1}$$

In the parametric case, this expression can be directly optimized; all that is needed is to insert the relevant expressions for $S_i(t)$. As in other cases, the likelihood function is maximized by differentiating the expression with respect to the parameters, and then setting these derivatives to 0.

In the non-parametric case, it becomes much more complicated. First, it is impossible to identify the survivor function at all time points. This corresponds to the problem that the Kaplan-Meier estimate cannot determine the tail of the distribution, when the largest time value corresponds to a censoring. For interval-censored data, this problem can occur at any time point. We can only determine the values of the survivor function at the interval endpoints, that is, the collection of L_i and R_i's, which we together will call x-values. In many cases, several consecutive of these x-values will show identical survivor function values, and thus the survivor function between them are given as the common value. When two consecutive values do not agree, we have an interval with positive estimated probability and we cannot determine where in the interval the probability mass lies. That is, we can only determine the total probability of that interval. This was realized already by Peto (1973), which also describes a procedure to select a subset of the intervals, which will contain all the probability mass. It is those intervals between the x-values, which have a lower endpoint among the L-observations and an upper endpoint among the R-observations. To emphasize that these make up only a subset of the intervals, they are denoted as intervals $(P, Q]$. The likelihood is formally the same as described in Eq. (19.1). When these intervals have been determined, estimation consists of optimising the likelihood subject the probabilities of these intervals being greater than or equal to 0. It is often the case that some of these intervals have zero estimated probability.

19.2 Proportional hazards models

This can be extended to the proportional hazards model, defined as the hazard being of the form

$$\lambda(t; z) = \lambda_0(t) \exp(\beta' z). \tag{19.2}$$

where z is vector of covariates with corresponding regression coefficients β and $\lambda_0(t)$ an arbitrary function describing the hazard as function of time. The regression parameters β is the interesting parameter, whereas the hazard function is a nuisance parameter. To apply this model to interval-censored data, we need to express this relation by means of the survivor function, for example as

$$S(t; z) = \exp\{-\Lambda_0(t) \exp(\beta' z)\}, \tag{19.3}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the integrated hazard function. The argument on selecting a subset of the intervals carries over without modification, because the argument does not request that the distributions are equal. Instead of using the probability parameters corresponding to each interval, we may use the contributions to the integrated hazards for each interval, say $\theta_j = \Lambda_0(Q_j) - \Lambda_0(P_j)$. This extends the proportional hazards model of Cox (1972), but the nice estimation methods of that paper cannot be used. The estimates can instead be found by generalizing the procedure from the non-parametric case.

19.3 Conditional proportional hazards

The proportional hazards model described above is very useful for finding the effect of covariates. However, it may still be relevant to extend the model, first of all, in its own right to obtain a more flexible model, when we think that the assumption of proportional hazard is not fulfilled and second as a means of goodness-of-fit checking the assumption of proportional hazards. One choice is the gamma frailty model, which specifies that conditional on the individual unobserved frailty, say Y the hazard has the form

$$\mu(t; z) = \mu_0(t) Y \exp(\beta' z). \tag{19.4}$$

As Y is unobserved, we have to assign a distribution to it and integrate out, to obtain the marginal distribution. Here we use the gamma distribution

$$f(y) = \delta^\delta y^{\delta-1} \exp(-\delta y)/\Gamma(\delta) \tag{19.5}$$

which is formulated to have a mean of 1. After integration we obtain the expression

$$S(t; z) = \{1 + \exp(\beta' z) M_0(t)/\delta\}^{-\delta}, \tag{19.6}$$

which can similarly be inserted into Eq. (19.1). This model will show converging hazards when Y has been integrated out. In that sense it is an extension of the

proportional hazards in only one direction. More details on the frailty model are described in Hougaard (2000).

This can be optimised in the same as for the proportional hazards model; there is just an additional parameter δ. This model can be compared to the proportional hazards model by the likelihood ratio test.

19.4 Conclusion

The proportional hazards has been suggested earlier for interval-censored data, but it seems to have been complicated to calculate the estimates and therefore, this has never become a standard model. However, estimation is not that difficult, so it is possible to apply this model. Indeed, it is not difficult to extend to the gamma frailty model, which is useful for setting the proportional hazards model in perspective.

References

1. Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B* **34**, 187–220.

2. Hougaard, P. (2000). *Analysis of multivariate survival data.*, Springer Verlag, New York.

3. Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics.* **22**, 86–91.

20

Optimal Maintenance Policies in Incomplete Repair Models

Waltraud Kahle

Otto-von-Guericke-University, Institute of Mathematical Stochastics, D-39016 Magdeburg, Germany

Abstract: We consider an incomplete repair model, that is, the impact of repair is not minimal as in the homogeneous Poisson process and not "as good as new" as in renewal processes but lies between these boundary cases. The repairs are assumed to impact the failure intensity following a virtual age process of the general form proposed by Kijima. In previous works field data from an industrial setting were used to fit several models. In most cases the estimated rate of occurency of failures was that of an underlying exponential distribution of the time between failures. In this paper it is shown that there exist maintanance schedules under which the failure behavior of the failure-repair process becomes a homogeneous Poisson process.

Keywords and phrases: Incomplete repair, Poisson process, renewal process, virtual age, hazard rate, optimal maintenance

20.1 Introduction

In this research, we are concerned with the statistical modeling of repairable systems. Our particular interest is the operation of electrical generating systems. As a repairable system, we assume the failure intensity at a point in time depends on the history of repairs. In the environment under investigation, it was observed that maintenance decisions were regularly carried out. We assume that such actions impacted the failure intensity. Specifically we assume that maintenance actions served to adjust the virtual age of the system in a Kijima type manner (Kijima et al., 1988, Kijima, 1989). Kijima proposed that the state of the machine just after repair can be described by its so-called virtual age which is smaller (younger) than the real age. In his framework, the rate of occurence of failures (ROCOF) depends on the virtual age of the system.

Kijima proposed two repair effect models. In his first model he assumed that repairs serve only to remove damage created in the last sojourn (a Kijima type I virtual age process). In his second model he assumed that the repair action could remove all damage accumulated up to that point in time (a Kijima type II virtual age process). That is, such repairs reset the virtual age of the unit to somewhere between that of a completely restored unit (good-as-new repair) and a minimally repaired unit, inclusively.

Further, we assume that the baseline failure intensity of the system follows a Weibull distribution

In Gasmi, Love, Kahle (2003) it was shown that the likelihood function can be developed from the general likelihood function for observation of point processes (Liptser and Shiryayev (1978)). Further, the likelihood ratio statistic can be used to find confidence estimates for the unknown parameters.

The numerical results for this data file are surprising: Under different assumptions about the repair actions (renewals, Kijima type I or II, mixture of Kijima type repairs and renewals in dependence on the time required for repair) a value for β was estimated approximately to be 1, see Gasmi, Love, Kahle (2003). That is, the failure intensity is more or less constant. But in this case the failure behavior does not depend on maintenance actions.

One of the reasons for this could be that for real systems, maintenance actions depend on the state of the system. In Kahle, Love (2003) it was assumed that each maintenance action has its own degree of repair which is assumed to be

$$\xi_k = 1 - \Phi(\log(r_k) - 2.4) \ ,$$

where ξ_k is the degree of repair after the kth failure or shut down, r_k is the repair time after the k-th sojourn and Φ is the distribution function of the standard normal distribution. The constant 2.4 is the estimated mean value of the log repair times. The estimated variance of the log repair times is about 1. It is easy to see that we get a degree of repair ξ_k of nearly 1.0 for very small repair times (which means that the age of the system after repair is the same as before the failure or shutdown) and a ξ_k of approximately 0.0 for long repair times (the system is perfectly repaired).

For this model we get the following estimates for the parameters of the baseline Weibull intensity:

$$\hat{\beta} = 2.305 \quad \hat{\alpha} = 134,645 \text{ min},$$

that is, the assumption that each maintenance action has its own degree of repair leads to an estimate of the shape parameter of $\hat{\beta} = 2.305$. This really increasing failure rate is in agreement with the experiences of maintenence engineers.

20.2 Optimal Maintenance as Time Scale Transformation

The results, mentioned in the previous sections, suggest, that in practice the engineers make a good maintenance policy, that is, they make repairs in dependence on the state of the system. The idea is that such a policy makes the apparent failure behavior of a system to be that of an exponential distribution. This is consistent with our data. In figure 20.1 are shown the cumulative distribution function of the operating time between failures together with the fitted CDF of an exponential distribution and the Q-Q plot (observed quantiles against the quantiles of the exponential model. These plots suggest reasonable agreement with the exponential model if we consider only the failure process and ignore all maintenance events.

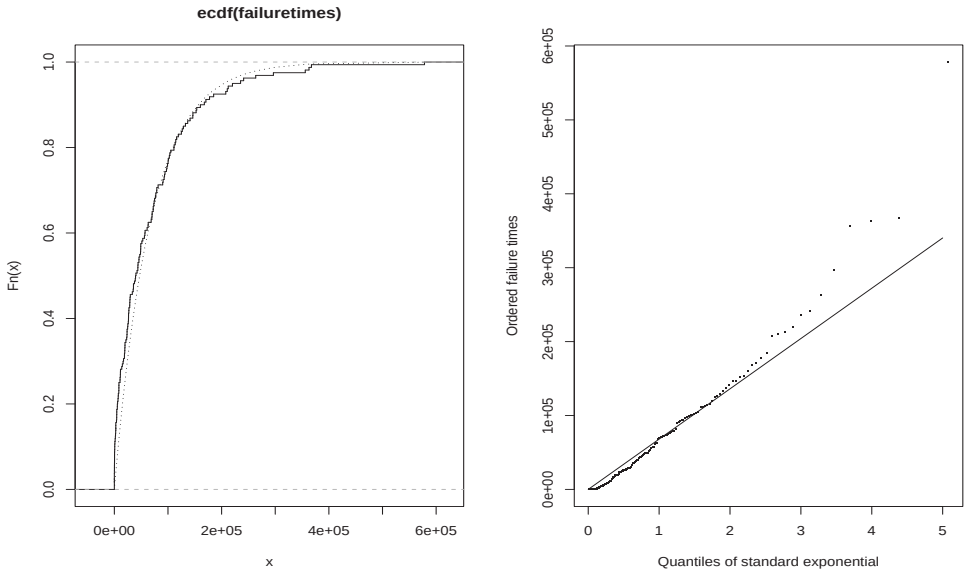


Figure 20.1: Operating time between failures: CDF and exponential Q-Q plot

Definition 20.1 *A maintenance policy is called* **failure rate optimal**, *if the state dependent preventive maintenance actions lead to a constant ROCOF of the failure process.*

Following an idea in Finkelstein (2000) we assume that by repair actions, the time scale is transformed by a function $W(t)$. Let $\Lambda_0(t)$ be the baseline

cumulative hazard function and let $\Lambda_1(t) = \Lambda_0(W(t))$ be the resulting hazard after a transformation of the time scale. For the Weibull hazard

$$\Lambda_0(t) = (t/\alpha)^\beta$$

and $W(t) = t^{1/\beta}$ we get

$$\Lambda_1(t) = \Lambda_0(t^{1/\beta}) = \frac{t}{\alpha^\beta} \ ,$$

that is, the hazard function of an exponential distribution with parameter $\lambda_1 = 1/\alpha^\beta$.

In practice we have repair actions at discrete time points, which leads to the question of the degrees of repair at these points. Let us consider two examples. In both examples we assume that after a failure the system is repaired minimally. Additionally, maintenance decisions were regularly carried out. We assume that maintenance actions served to adjust the virtual age of the system in a Kijima type manner.

Example 1: Assume that the distances between maintenance actions are constant and all repair actions follow the Kijima type I repair process. Let $t_1, t_2, \ldots$ be the time points of maintenance actions and $\Delta = t_k - t_{k-1},\ k = 1, 2, \ldots$, where $t_0 = 0$, be the constant distance between maintenances. Then it is possible to find a discrete time transformation which consists of different degrees of repair. Let the sequence of degrees be

$$\xi_k = \frac{k^{1/\beta} - (k-1)^{1/\beta}}{\Delta^{1-1/\beta}} \ .$$

Then the virtual age v_n of the system at time $t_n = n \cdot \Delta$ can be found to be

$$v_n \ = \ \Delta \sum_{k=1}^{n} \xi_k = \Delta \sum_{k=1}^{n} \frac{k^{1/\beta} - (k-1)^{1/\beta}}{\Delta^{1-1/\beta}} = (n \cdot \Delta)^{1/\beta} \ .$$

Example 2: Again we assume that the distances between maintenance actions are constant, but now we consider the Kijima type II repair process. In this case the appropriate sequence of degrees of repair is

$$\xi_k = \frac{k^{1/\beta}}{(k-1)^{1/\beta} + \Delta^{1-1/\beta}} \ .$$

In both cases the sequence is decreasing, that is, with increasing time the repairs must become better.

In figure 20.2 are shown the cumulative hazard functions for an Weibull process without maintenance (solid line) and for maintenance actions every $\Delta = .1$ time units (broken line). For this, a Weibull process with parameters $\alpha = 1$ and $\beta = 2.5$ and 30 failures was simulated. The difference $\Delta = .1$ between maintenance actions is relatively small, and the empirical cumulative hazard function of the process with preventive maintenance is closed to that of a Poisson process. The dotted line shows the theoretical cumulative hazard function of an homogeneous Poisson process.

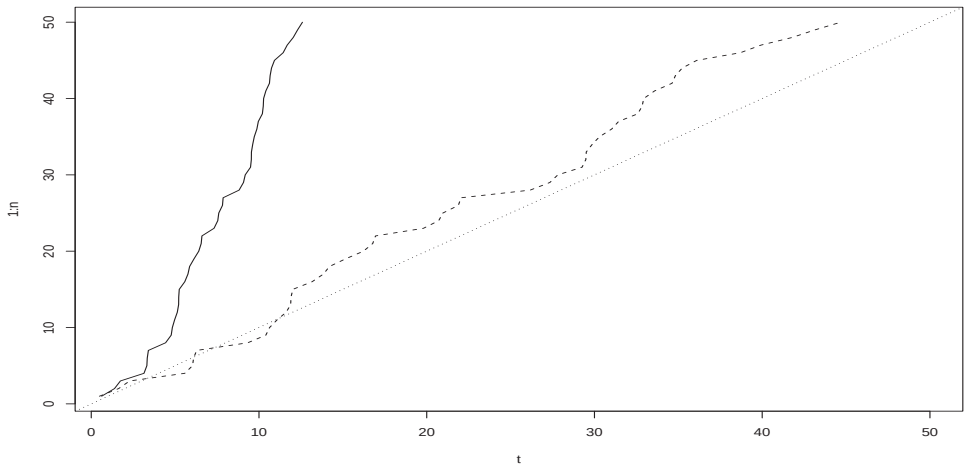


Figure 20.2: Weibull process without and with preventive maintenance actions

There are many other possibilities for finding failure rate optimal maintenance policies. One other very simple policy is to consider constant degrees of repair. It is easy to see that in this case the repair actions must take place more often with increasing time.

References

1. Finkelstein, M. S. (2000). Modeling a Process of Non-Ideal Repair. In *Recent Advances in Reliability Theory* (Eds. N. Limnios, M. Nikulin), pp. 41–53, Birkhhauser, Series Statistics for Industry and Technology.

2. Gasmi, S., C.E. Love, and Kahle, W. (2003). A General Repair/Proportional Hazard Framework to Model Complex Repairable Systems, *IEEE*

Trans. on Reliability, **52**, 26–32.

3. Kahle, W. and C. E. Love (2003). Modeling the Influence of Maintenance Actions. In *Mathematical and Statistical Methods in Reliability* (Eds. B.H. Lindquist, K. A. Doksum), pp. 387–400, World Scientific Publishing Co., Series on Quality, Reliability and Engeneering Statistics.

4. Kijima, M. (1989). Some results for repairable systems with general repair, *Journal of Applied Probability*, **26**, 89-102 (1989).

5. Kijima, M., H. Morimura, and Y. Suzuki (1988). Periodical Replacement Problem without Assuming Minimal Repair, *European Journal of Operational Research*, **37**, 194–203.

6. Last, G. and R. Szekli (1998). Stochastic Comparison of Repairable Systems by Coupling, *Journal of Applied Probability*, **35**, 348-70.

7. Liptser, R.S. and A.N. Shiryayev (1978). *Statistics of Random Processes, vol II*, Springer, New York.

21

Binary Regression in Truncated Samples, with Application to Comparing Dietary Instruments in a Large Prospective Study

Victor Kipnis, Douglas Midthune, Laurence S. Freedman and Raymond J. Carroll

Biometry Research Group, National Cancer Institute, EPN-3131, 6130 Executive Blvd., MSC 7354, Bethesda, MD 20892-7354, U.S.A.
Biometry Research Group, National Cancer Institute, EPN-3131, 6130 Executive Blvd., MSC 7354, Bethesda, MD 20892-7354, U.S.A.
Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer 52161, Israel
Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, U.S.A.

21.1 Introduction

This talk is concerned with two important questions in nutritional epidemiology. The first is biological: is there a relationship between dietary fat intake and breast cancer? The second is methodological: how do various dietary assessment instruments compare in their power to detect diet-disease relationships?

In studying diet-disease relationships, the dietary assessment instrument most widely used is the Food Frequency Questionnaire (FFQ) which is relatively cheap to administer, and hence is practical to use in large-scale prospective studies. Recently, a series of cohort prospective studies with FFQ has found little evidence of a fat-breast cancer relationship.

There are other dietary assessment instruments, however, including multiple-day food records. These instruments are more expensive to administer, but are often thought to be better measures of dietary intake than the FFQ. Recently, Bingham et al. (2003) reported the results of a comparison of two instruments, a FFQ and a 7-day diary, both completed by a cohort of 13,070 women. They found that a statistically significant positive association between fat intake and breast cancer incidence could be demonstrated using the 7-day diary, but not using the FFQ. This study suggested that the 7-day diary, being associated with less error than the FFQ, may be more efficient than the FFQ at detecting this diet-disease relationship. However, the study was based on a relatively small number of breast cancer cases (168).

We undertook a similar comparison within the control group of the Dietary Modification (DM) arm of the Women's Health Initiative (WHI) Clinical Trial. This is a randomized controlled trial of a low-fat diet that is high in fruits, vegetables, and grains. General eligibility criteria for the trial are provided in detail elsewhere (Hays et al., 2003). Women who participated in this trial completed both a FFQ and a 4-day food record (FR) on entry, allowing a comparison between these instruments, similar to the comparison of Bingham et al. To achieve more power in comparing the dietary intervention and control groups, women were excluded from the trial if they reported consuming a diet with less than 32% energy from fat, as estimated by the FFQ.

The control group comprised approximately $30,000$ women who received general advice on diet and health, but no intensive dietary counseling. At the time of the analysis, after a median follow up of 7 years, 603 of these women had been diagnosed with invasive breast cancer. Staff at the WHI Clinical Coordinating Center frequency matched 2 women with no diagnosis of breast cancer for every case, matched on age category (50-59, 60-69, 70-79), clinic, and length of follow-up (+/- 12 months) resulting in a sample of 603 cases and 1206 controls. The matching was done because the cost of analyzing the food records of all women in the control group would have been prohibitive.

With this as background, this talk has two main goals. The first goal is to describe and compare two methods for estimating relative risks in the entire (non-truncated) population using data from a truncated sample, both for the FFQ (on which truncation is based) and for the FR. One method uses a full (prospective) likelihood. The other is an approximate method that posits a risk model in which the truncation variable is included, and then employs a new residualization method. We will demonstrate that the approximate method gives estimates with very small bias and variances that are substantially smaller than those of the full likelihood method.

The second goal of this talk is to describe a simple approximate methodology that allows the comparison of the instruments, again for the entire, rather than the truncated, population. The key point is that we want to compare the relative local power of instruments *that would be obtained in a non-truncated prospective study*, using data from the truncated sample. To do so, we develop estimates of the standard errors of the estimators that would have been obtained in a (hypothetical) non-truncated sample. We show that under some fairly mild approximations, one can obtain an approximate comparison of the instruments without having to model the often high-dimensional distribution of all the covariates.

21.2 Models for Risk Analysis in Truncated Samples

Let Y be a binary response, let F be the scalar risk factor of primary importance, and let X be all the other covariates. The risk model of interest is the logistic regression model

$$P(Y = 1|F, X) = H\left(\beta_0 + F\beta_1 + X'\beta_2\right), \tag{21.1}$$

where $H(\cdot)$ is the logistic distribution function. Interest is in the log relative risk β_1. Because of biased sampling, we observe only those subjects with $C > c_{\text{trun}}$, where C is a variable that is related to F and X, and c_{trun} is the truncation cutoff value.

21.2.1 A prospective likelihood approach

Suppose that C is normally distributed within the cases and controls as follows:

$$[C|Y = y, F, X] = Normal(\alpha_{0y} + \alpha'_{1y}F + \alpha'_{2y}X, \sigma_y^2), \tag{21.2}$$

for $y = 0, 1$. The probability of being in the truncated sample given disease status is

$$P(C > c_{\text{trun}}|Y = y, F, X) \quad = \quad 1 - \Phi\left\{\frac{c_{\text{trun}} - \alpha_{0y} - \alpha'_{1y}F - \alpha'_{2y}X}{\sigma_y}\right\} \tag{21.3}$$

Defining $\mathcal{A} = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \sigma_0, \alpha_{01}, \alpha_{11}, \alpha_{21}, \sigma_1)'$ and

$$S(F, X, \mathcal{A}) = \log\left\{\frac{P(C > c_{\text{trun}}|Y = 1, F, X)}{P(C > c_{\text{trun}}|Y = 0, F, X)}\right\}, \tag{21.4}$$

it follows easily that the truncated sample has the risk model

$$P(Y = 1|X, F, C > c_{\text{trun}}) = H\left\{\beta_0 + \beta'_1F + \beta'_2X + S(F, X, \mathcal{A})\right\}. \tag{21.5}$$

Using equation (21.2), the density of C given disease status in the truncated sample is

$$\begin{aligned} f_{C,\text{trun}}(c|Y = y, X, F, C > c_{\text{trun}}) = \frac{1}{\sigma_y}\phi\left\{\frac{c - \alpha_{0y} - \alpha'_{1y}F - \alpha'_{2y}X}{\sigma_y}\right\} \\ \times\left[1 - \Phi\left\{\frac{c_{\text{trun}} - \alpha_{0y} - \alpha'_{1y}F - \alpha'_{2y}X}{\sigma_y}\right\}\right]^{-1}, \end{aligned} \tag{21.6}$$

and the joint density of (Y, C) in the truncated sample is

$$\begin{aligned} f_{Y,C,\text{trun}}(y, c|X, F, C > c_{\text{trun}}) = f_{C,\text{trun}}(c|Y = y, X, F, C > c_{\text{trun}}) \\ \times\left[H\left\{\beta_0 + \beta'_1F + \beta'_2X + S(F, X, \mathcal{A})\right\}\right]^y \\ \times\left[1 - H\left\{\beta_0 + \beta'_1F + \beta'_2X + S(F, X, \mathcal{A})\right\}\right]^{1-y}. \end{aligned} \tag{21.7}$$

Equation (21.7) is the *prospective* density function conditional on (X, F) and the population formed by the truncated sample, in terms of the parameters $(\beta_0, \beta_1, \beta_2)$ and $\mathcal{A}$. Consistent and asymptotically normal estimates of these parameters are obtained by maximizing the prospective likelihood based upon (21.7), even ignoring the nested case-control sampling scheme.

21.2.2 A more powerful approximate approach

In simulations and in the analysis of the WHI controls study, we found that the above likelihood approach gave risk estimates with too large variability due to poorly estimated offset $S(F, X, \mathcal{A})$. Therefore we developed an alternative methodology.

Assume that the risk model for Y given (F, X, C) is linear logistic in (F, X, C). Note that any regression model which includes the truncation variable C as a covariate will be immune from any bias arising from the truncated sample. Inclusion of C in the model, however, will change the regression parameters of the other covariates (F and X in our case), if C is correlated with them. To overcome this latter problem, we define $R = C - E(C|F, X)$, the residual of the regression of C on (F, X). If $E(C|F, X)$ is linear in (F, X), then Y is linear logistic in (F, X, R), and R is uncorrelated with the other covariates. Furthermore, since by conditioning on (F, X, R) we are conditioning on C, the model is still immune from bias due to truncation.

Under our assumptions, Y approximately follows the linear logistic risk model in (F, X, R):

$$P(Y = 1|F, X, R) \approx H\left(\beta_0 + F\beta_1 + X'\beta_2 + R\beta_3\right).$$

In order to implement this method, we must estimate $E(C|F, X)$ within the truncated sample. In the present nested case-control study, assuming a rare disease, this can be handled approximately under assumption (21.2) using the control data only.

21.3 Comparison of Instruments

Let π_0 be the marginal probability of disease in the population. Let σ_F^2 be the variance of F in the population, let its covariance with X be Σ_{FX} and let the covariance matrix of X be Σ_{XX}. Then

$$n^{1/2}(\widehat{\beta}_1 - \beta_1) \approx \text{Normal}\left[0, \sigma^2_{\beta,1} = \left\{\pi_0(1-\pi_0)(\sigma^2_{\rm F} - \Sigma_{\rm FX}\Sigma^{-1}_{\rm XX}\Sigma'_{\rm FX})\right\}^{-1}\right], \quad (21.8)$$

Local power is determined by the noncentrality parameter, namely

$$\Theta = n^{1/2}\beta_1/\sigma_{\beta,1}. \quad (21.9)$$

Let Θ_{FFQ} and Θ_{FR} be the noncentrality parameters for the FFQ and FR, respectively, and let $\widehat{\Theta}_{FFQ}$ and $\widehat{\Theta}_{FR}$ be estimates of Θ_{FFQ} and Θ_{FR}. Inference about the difference in the local power of the two instruments can be based on $\kappa = \Theta_{FR} - \Theta_{FFQ}$ and its estimate $\widehat{\kappa} = \widehat{\Theta}_{FR} - \widehat{\Theta}_{FFQ}$, using bootstrap methods to estimate the standard error of $\widehat{\kappa}$.

Suppose that the bivariate pair (C, F) are jointly normally distributed given X, with means linear in X and a constant covariance matrix. This implies the models

$$[F|C, X] = \alpha_0 + \alpha_1 C + \alpha_2' X + \epsilon; \tag{21.10}$$

$$[C|X] = \gamma_0 + \gamma_1' X + \eta, \tag{21.11}$$

where ϵ has mean zero and variance σ_ϵ^2, while η has mean zero and variance σ_η^2. It is easily shown that

$$\sigma_{\beta,1}^2 = \left\{\pi_0(1-\pi_0)(\sigma_\epsilon^2 + \alpha_1^2\sigma_\eta^2)\right\}^{-1}, \tag{21.12}$$

In order to implement (21.12), we need to estimate σ_ϵ^2, α_1 and σ_η^2. This can be done as follows. First, estimate σ_ϵ^2 and α_1 by regressing F on (C, X) in the truncated sample. Then, estimate σ_η^2 by maximum likelihood using the likelihood based on model (21.11). In the WHI controls study, estimation is complicated by the nested case-control sampling scheme. Under a rare disease assumption, one can estimate $\sigma_{\beta,1,\text{approx}}^2$ by fitting (21.10) and (21.11) among the controls only.

21.4 Analysis of the WHI Data

In the present study, the exposure of interest, F, is the logarithm of total fat, saturated fat, polyunsaturated fat or monounsaturated fat, as estimated from the FFQ or FR. The vector of other risk factors, X, consists of the following variables: logarithm of energy intake, duration of follow-up, age at entry to study (in 5-year age groups), clinical center region (North-East, South, Mid-West, West), hormone use (never, ever), family history (missing, no, yes), and breast biopsy (missing no, yes). The truncation variable C is the logarithm of FFQ-reported percent calories from fat.

For the approximate method, when using the FR to assess diet, total fat, polyunsaturated fat and monounsaturated fat are statistically significant risk factors for breast cancer incidence, with estimated log relative risks of 0.78 (s.e.=0.27), 0.51 (s.e.=0.18) and 0.71 (s.e.=0.23), respectively, while saturated fat, with an estimated log relative risk of 0.31 (s.e.=0.19), is not statistically significant. When using the FFQ to assess diet, none of the types of dietary fat are statistically significant risk factors, and the estimated log relative risks are generally much smaller than those estimated using the FR.

When we compare the local power of the FR and FFQ using $\widehat{\kappa}$, the estimated difference in noncentrality parameters in a hypothetical non-truncated sample, tests do not reach the formal 0.05 level of significance for any of the types of dietary fat; the tests for total fat ($\widehat{\kappa} = 2.07$, s.e.= 1.18) and polyunsaturated fat ($\widehat{\kappa} = 2.25$, s.e.= 1.26), however, are close to significant, with p-values of 0.08 and 0.07, respectively, strongly suggesting that the FR has greater local power to detect these diet-disease relationships.

Three of the four types of dietary fat estimated by FR were statistically significant risk factors when we used the approximate method to estimate regression parameters, while none were significant when we used the (prospective) likelihood method. The standard errors of $\widehat{\beta}_1$ for the likelihood method, moreover, were 30% to 50% larger than those for the approximate method, which is similar to what we saw in simulations.

References

1. Bingham, S. A., Luben, R., Welch, A., Wareham, N., Khaw, K. T. and Day, N. (2003). Are imprecise methods obscuring a relation between fat and breast cancer? *Lancet*, 362, 212-214.

2. Hays, J., Hunt, J. R., Hubbell, F. A., Anderson, G. L., Limacher, M., Allen, C. and Rossouw, J. (2003). The Women's Health Initiative: recruitment, methods and results. *Annals of Epidemiology*, 13, S18-S77.

22

Comparison of Sequential Experiments for Estimating the Number of Classes in a Population

Subrata Kundu and Tapan K. Nayak

Department of Statistics, George Washington University, Washington, DC 20052

Abstract: This paper deals with stopping rules in sequential sampling from a population with an unknown number of classes or species. It is assumed that the members of the population are selected one at a time and all classes are equally likely to occur in each selection. Blackwell's criterion for a "more informative experiment" is used to compare all stopping rules and derive certain complete class results. It is also shown that only non-randomized stopping rules may yield minimal sufficient statistics that are complete.

Keywords and phrases: Admissibility, capture-recapture, completeness, stopping rule, sufficiency

22.1 Introduction

Consider a population consisting of an unknown number, ν ($\nu \geq 1$), of classes. Each member of the population belongs to one of the ν classes and conversely each class has at least one representative in the population. In practical applications, a class may represent a biological species, a word in a vocabulary, an error in a software code, a demographic category, a genotype etc. Any specific class is discovered when the first member of that class is observed. Suppose we randomly select the population members sequentially, one at a time, and with replacement if the population is finite. For each selected unit we record if it belongs to a new class or not. Let $X_j = 1$ if a new class is discovered in the jth selection, and $X_j = 0$ otherwise. Then, the data generated by n selections can be represented by $X_1, X_2, \ldots X_n$. For n selections, let $x^{(n)} = (x_1, x_2, \ldots, x_n)$ denote the observed data, and let $R_n(x^{(n)}) = \sum_{j=1}^{n} x_j$ and $M_n(x^{(n)}) = n - R_n(x^{(n)})$ denote the number of discovered classes and the number of repeat events, respectively. The outcome of a sequence of selections

can also be viewed as a path in a two dimensional lattice of points with non-negative integer co-ordinates. All paths start at the origin $(0, 0)$, and at the jth step move one unit to the right if $x_j = 1$, and one unit up if $x_j = 0$.

Estimation of the number of classes ν has been discussed in a variety of forms by many authors, mostly assuming that all classes are equally likely to occur in each selection. One goal of this paper is to extend the investigation of completeness for randomized stopping rules. However, much of this paper is devoted to comparison of stopping rules and some related results. We shall assume that all classes are equally likely to occur in each selection, which is a common assumption for estimation of ν. This assumption is important for unbiased estimation of ν; without any assumption about the class probabilities, ν cannot be estimated unbiasedly even when ν has an upper bound (cf., Christman and Nayak, 1994).

A stopping rule ϕ is a sequence of functions $\phi_0, \phi_1(x^{(1)}), \phi_2(x^{(2)}), \ldots$, where $\phi_j(x^{(j)})$ is the conditional probability of stopping sampling given that j selections have been made with observed outcomes $x_1, x_2, \ldots, x_j$. As the first selection always discovers a new class, we shall take $\phi_0 = \phi_1 = 0$. Let N denote the sample size, which may be a random variable. Then, we have: $\phi_j(x^{(j)}) = P(N = j | N \geq j, X^{(j)} = x^{(j)})$. We shall consider only those stopping rules ϕ (or equivalently ψ) for which sampling stops with probability 1.

In this paper we present relevant distribution theory and sufficient statistics. We find that $\{R_n\}$ is a minimal sufficient sequence and it is transitive. This implies that the stopping rules based on (R_n, M_n) form an essentially complete class. Then we compare these stopping rules using Blackwell's (1951) criterion for comparing experiments. We show that a 'more informative' stopping rule also costs more in terms of expected sample size. We consider stopping rules with bounded average sample size and present certain complete class results. Some of our results are similar to those of Kusama and Koyama (2000) for sequential binomial experiments. Our context, however, differs from sequential binomial sampling in two significant ways: Unlike in binomial sampling, our sample space depends on the unknown parameter ν, and successive observations in our case are not independent. Finally, we investigate completeness and show that a closed complete stopping rule must be non-randomized.

22.2 Sufficient Statistics.

In this section we present some distributional results and minimal sufficient statistics.

Definition 22.2.1 *For each $n \geq 1$, let T_n be a function of $X^{(n)}$. Then, $\{T_n, n \geq 1\}$ is said to be a sufficient sequence for ν if for each n, T_n is a sufficient statistic for ν based on $X^{(n)}$.*

It is easy to see that $R_n(x^{(n)})$ is a minimal sufficient statistic. Thus, in our context $\{R_n\}$ is a minimal sufficient sequence. Then, from Blackwell (1947) it follows that for any given stopping rule $\phi, (N, R_N)$ or equivalently $(R_N, M_N = N - R_N)$ is sufficient for ν, i.e., the conditional distribution of $X^{(N)}$ given (N, R_N) is independent of ν. We shall take

$$G = \{\alpha = (r, m) : r \text{ and } m \text{ are non-negative integers}\}, \tag{22.2.1}$$

which is the two-dimensional lattice with non-negative integer points, as a common sample space of (R_N, M_N) for all ϕ.

For a sequential experiment, a decision rule δ consists of a sequence of decision functions $\delta_0, \delta_1(x^{(1)}), \delta_2(x^{(2)}), \cdots$, where $\delta_j(x^{(j)})$ is the action to be taken if sampling stops after observing $x^{(j)}$. Let $L(\nu, a)$ be the loss from taking action a when the true parameter value is ν, and let $c_j(\nu, x^{(j)})$ denote the sampling cost of taking observations $x_1, \cdots, x_j$ and stopping. The total cost of stopping at $x^{(j)}$ and then taking action $\delta(x^{(j)})$ is $L(\nu, \delta(x^{(j)})) + c_j(\nu, x^{(j)})$. The risk function, to be denoted by $R(\nu, (\phi, \delta))$, is the expected total cost from using the stopping rule ϕ and the decision rule δ. For any given stopping rule ϕ, the decision rules based on a sufficient sequence form an essentially complete class (see, Ferguson (1967), Sec. 7.3), i.e., given any decision rule δ, there exists a decision rule δ^0 based only on a sufficient sequence $\{T_j\}$ such that $R(\nu, (\phi, \delta^0)) \leq R(\nu(\phi, \delta))$ for all $\nu \geq 1$. Furthermore, if the loss function is convex any decision rule that is not based on $\{T_j\}$ can be improved upon by Rao-Blackwellization.

In sequential decision theory, however, it is also known that for a given pair of stopping and decision rules (ϕ, δ) there need not exist a pair (ϕ^0, δ^0), where both ϕ^0 and δ^0 are based on a sufficient sequence $\{T_j\}$ and $R(\nu, (\phi^0, \delta^0)) \leq R(\nu(\phi, \delta))$ for all $\nu \geq 1$. Bahadur (1954) showed that a positive assertion in this direction can be made if $\{T_j\}$ is transitive:

Definition 22.2.2 *A sufficient sequence $\{T_j\}$ is said to be transitive if for every $j \geq 1$, and all bounded integrable functions g,*

$$E[g(X^{(j)})|T_j, T_{j+1}] = E[g(X^{(j)})|T_j].$$

In our case the sequence $\{R_j\}$ is a transitive sufficient sequence. From Bahadur's (1954) general results we now have the following:

Theorem 22.2.1 *For any given stopping rule ϕ, there exists a stopping rule ϕ^0 based on $\{R_j\}$ such that the distribution of (N, R_N) under ϕ is the same as the distribution of (N, R_N) under ϕ^0 for all $\nu \geq 1$.*

Theorem 22.2.2 *If the sampling cost $c_j(\nu, x^{(j)})$ depends on $x^{(j)}$ only through $R_j(x^{(j)})$, the class of rules (ϕ^0, δ^0) based on $\{R_j\}$ is essentially complete, that is, given any rule (ϕ, δ) there exists a rule (ϕ^0, δ^0) based on $\{R_j\}$ such that $R(\nu, (\phi^0, \delta^0)) \leq R(\nu(\phi, \delta))$ for all $\nu \geq 1$.*

The assumption about sampling cost in Theorem 22.2.2 is satisfied in most practical applications; often $c_j(\nu, x^{(j)})$ is a function only of the number of selections j, e.g., $c_j(\nu, x^{(j)}) = kj$. In view of Theorem 22.2.2 we shall consider only the stopping rules in the class $\mathcal{C}_0 = \{\phi : \phi_j(x^{(j)}) \text{ depends on } x^{(j)} \text{ only through } R_j \text{ and } P_\nu(N < \infty|\phi) = 1\}$.

22.3 Comparison of Stopping Rules.

In this section we shall compare the stopping rules in $\mathcal{C}_0$ using Blackwell's (1951) ideas for comparing experiments. In general, an experiment is defined by a sample space $\mathcal{X}$, a σ-algebra, and a family of probability distributions on $\mathcal{X}$ indexed by an unknown parameter. In our application, an experiment is equivalent to a stopping rule. Applying sufficiency we shall assume that only the number of discoveries (R) and repeat events (M) at the stopping time are observed and hence we shall take G (defined in (22.2.1)) as a common sample space of all of our experiments. We now state Blackwell's (1951) criterion for comparing two stopping rules (or sampling plans).

Definition 22.3.1 *A stopping rule ϕ_1 is said to be sufficient for (more informative than) another stopping rule ϕ_2 if there exists a stochastic kernel z, i.e., a probability measure $z(.|\alpha)$ on G for each $\alpha \in G$, such that*

$$P_\nu(\beta|\phi_2) = \sum_{\alpha \in G} z(\beta|\alpha) P_\nu(\alpha|\phi_1) \quad \forall \quad \nu \geq 1, \beta \in G.$$

We shall write $\phi_2 \preceq \phi_1$ when ϕ_1 is sufficient for ϕ_2. If $\phi_2 \preceq \phi_1$ and $\phi_1 \preceq \phi_2$, ϕ_1 and ϕ_2 are said to be equivalent, in which case we shall write $\phi \sim \phi_2$. If $\phi_2 \preceq \phi_1$ but $\phi \not\sim \phi_2$, ϕ_1 is said to be more informative that ϕ_2, to be written as $\phi_2 \prec \phi_1$.

From Blackwell (1953) it follows that if ϕ_1 is sufficient for ϕ_2, then for every loss function $L(\nu, a)$ (not accounting for the cost of sampling), and decision rule δ based on ϕ_2, there exists a decision rule δ_* based on ϕ_1 such that $E_{\nu,\phi_2} L(\nu, \delta(\alpha)) \leq E_{\nu,\phi_1} L(\nu, \delta_*(\alpha))$ for all $\nu \geq 1$. Definition 22.3.1 can be used naturally to define admissible stopping rules: For a given class of stopping rules $\mathcal{C}$, $\phi \in \mathcal{C}$ is said to be admissible in $\mathcal{C}$ if there does not exist any $\phi^* \in \mathcal{C}$ such that $\phi \prec \phi^*$. By Definition 22.3.1, if ϕ_1 is sufficient for ϕ_2, then the sampling distribution under ϕ_2 can be achieved by using a random transformation (not involving ν) after observing the outcome under ϕ_1. We next present a condition that those random transformations must satisfy.

Lemma 22.3.1 *Suppose ϕ_1 is sufficient for ϕ_2, i.e., there exists a kernel z, such that for each $\beta \in G$, $P_\nu(\beta|\phi_2) = \sum_{\alpha \in G} z(\beta|\alpha) P_\nu(\alpha|\phi_1)$ for all $\nu \geq 1$. For any given $\delta \in G$ if $P_\nu(\delta|\phi_1) > 0$ for some $\nu \geq 1$, then $z\,(L(\delta)|\delta) = 1$, where $L(\delta) = \{\alpha \in G : r(\alpha) \leq r(\delta), m(\alpha) \leq m(\delta)\}$ is the lower left quadrant of δ.*

In the following we show that more informativeness comes only at a cost of higher average sample size (Theorem 22.3.1), and that any stopping rule can be modified to obtain a more informative rule if the average sample size is allowed to increase, even by an arbitrarily small amount (Theorem 22.3.2).

Theorem 22.3.1 *If ϕ_1 is sufficient for ϕ_2, then $E_\nu[N|\phi_1] \geq E_\nu[N|\phi_2]$. Further, if $\phi_2 \prec \phi_1$, then the inequality is strict.*

Theorem 22.3.2 *For any sampling plan $\phi \in \mathcal{C}_0$ and any $\epsilon > 0$, there exists a sampling plan ϕ^* such that $\phi \prec \phi^*$ and $0 < E_\nu[N|\phi^*] - E_\nu[N|\phi] < \epsilon$ for all $\nu > 0$.*

22.4 Plans With Bounded Average Sample Size.

In this section we shall consider the following subclasses of $\mathcal{C}_0$: $S_k = \{\phi \in \mathcal{C}_0 | \sup_\nu E_\nu(N|\phi) \leq k\}$, $S_{0,k} = \{\phi \in S_k | \sup_\nu E_\nu(N|\phi) = k\}$ and $\mathcal{A}_k = \{\phi \in S_k | \phi \text{ is admissible in } S_k\}$.

Our main goal is to discuss admissible plans within S_k. In particular, we state a sequence of results to finally show that $\mathcal{A}_k$ is minimal complete but not all plans in $S_{0,k}$ are admissible.

Theorem 22.4.1 $\mathcal{A}_k \subset S_{0,k}$.

Theorem 22.4.2 *If $E_\nu(N|\phi) = k$ for some ν, then ϕ is admissible in S_k. In particular, the plan with fixed sample size k is admissible in S_k.*

Theorem 22.4.3 *$\mathcal{A}_k$ is minimal complete in S_k. Then, by Theorem 4.1, $S_{0,k}$ is complete in S_k.*

Theorem 22.4.4 *$S_{0,k}$ is not minimal complete in S_k.*

22.5 Completeness.

A stopping rule ϕ is said to be complete if for any function f defined on G, $E_\nu[f(\alpha)] = 0$ for all $\nu \geq 1$ implies $f(\alpha) = 0$ for all $\alpha \in G$. Completeness is a helpful property for unbiased estimation and hypotheses testing. In this section we characterize completeness of closed stopping rules in $\mathcal{C}_0$. A stopping rule ϕ is closed if $P_\nu(N < \infty|\phi) = \sum_{\alpha \in G} P_\nu(\alpha|\phi) = 1$ for all $\nu \geq 1$.

Theorem 22.5.1 *A closed stopping rule $\phi \in \mathcal{C}_0$ is complete only if ϕ is a non-randomized rule, i.e., for each $\alpha \in G, \phi(\alpha)$ is either 0 or 1.*

The last theorem shows that complete stopping rules belong to the subclass of non-randomized rules. Christman and Nayak (1994) investigated that subclass and characterized the complete rules in it. Thus, Theorem 22.5.1 together with their results give a full account of complete rules.

Desire for complete stopping rules lead us to non-randomized rules. However, we note that non-randomized rules may not be maximally informative within a given class of stopping rules. More generally, we can show that for any k, the non-randomized rules in S_k do not form an essentially complete class (within S_k).

References

1. Bahadur, R. R. (1954). Sufficiency and statistical decision functions, *Ann. Math. Statist.* **25**, 423-462.

2. Blackwell, D. (1951). Comparison of Experiments, *Proc. Second Berkeley Sympos. Math. Statist. Probab.*, 93-102.

3. Blackwell, D. (1953). Equivalent comparison of experiments, *Ann. Math. Statist.* **24**, 265-272.

4. Christman, M., and Nayak, T. K. (1994). Sequential unbiased estimation of the number of classes in a population, *Statistica Sinica* **4**, 335-352.

5. Ferguson, T. S. (1967). *Mathematical Statistics : A Decision Theoretic Approach*, Academic Press (New York).

6. Kusama, T., and Koyama, N. (2000). Complete classes in the comparison of sequential binomial experiments, *Stat. and Dec.* **18**, 27-34.

7. Nayak, T. K. (1992). On statistical analysis of a sample from a population of unknown species, *J. Statist. Plann. Inference* **31**, 187-198

23

Estimation of Rescaled Distribution for Semi-parametric Goodness of Fit

H. Läuter†, M. Nikulin‡, V. Solev††

†*University of Potsdam,*
‡*Université Victor Segalen, Bordeaux II,*
††*Steklov Institute of Mathematics, Saint Petersburg*

23.1 Extended Abstract

Let $Z = (X, Y)$ be a random vector, P_Z be the distribution of Z, P^X, P^Y be the distributions of X, Y correspondingly . We consider the product measure

$$P_Z^{\times} = P_X \times P_Y.$$

We assume that the measure P_Z is absolutely continuous with respect to the measure $P_Z^{\times}$ and set

$$p(x, y) = \frac{dP_Z}{dP_Z^{\times}}(x, y).$$

We suppose that there exists the density function $f(x, y)$ of the distribution P_Z with respect to the Lebesgue measure. The density function f can be represented in the form:

$$f(x, y) = p(x, y) f_X(x) f_Y(x), \tag{1}$$

where f_X, f_Y are the density functions of the random variables X, Y.

We denote by F_X, F_Y the distribution functions of the random variables X, Y and put

$$q(x, y) = p\left(F_X^{-1}(x), F_Y^{-1}(y)\right), \tag{2}$$

where F^{-1} is the notation for the inverse function. The function $q(x, y)$ is density function on $[0, 1] \times [0, 1]$. It is the density function of the rescaled random vector

$$Z^* = (X^*, Y^*) = \left(F_X(X), F_Y(Y)\right). \tag{3}$$

It is clear that the random variables X^*, Y^* are uniformly distributed on $[0, 1]$. The density function $q(x, y)$ is called copula density and is responsible for the type of dependence of the coordinates of the vector Z.

So, we have

$$f(x,y) = q\left(F_X(x), F_Y(y)\right) f_X(x) f_Y(y), \tag{4}$$

Let $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ be i.i.d. random vectors with common distribution P_Z, P_n^X and P_n^Y be the empirical measures constructed on samples $X_1 \ldots, X_n$ and $Y_1 \ldots, Y_n$ correspondingly,

$$P_n^X\{A\} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_A(X_j), \quad P_n^Y\{A\} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_A(Y_j),$$

and F_n^X, F_n^Y be the corresponding empirical distribution functions. Denote

$$Z_j^* = \big(F^X(X_j), F^Y(Y_j)\big), \quad Z_{n,\,j}^* = \big(F_n^X(X_j), F_n^Y(Y_j)\big). \tag{5}$$

and set

$$P_n^{Z^*} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_A(Z_j^*) \;\; \mathcal{P}_n^{Z^*} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_A\big(Z_{n,\,j}^*\big).$$

Suppose that we observe the sample $Z_1, \ldots, Z_n$. If the distributions of X and Y are unknown, then it is impossible to construct the empirical measure $P_n^{Z^*}$ on observations $Z_1, \ldots, Z_n$. There are many reasons to think (see in details Rüschendorf L. (1976)) that the observable "empirical" measure $\mathcal{P}_n^{Z^*}$ is close (in a certain sense) for large n to $P_n^{Z^*}$.

For a metric space $([0,1]\times[0,1], r)$ we consider Kantorovich metric $\kappa_r(\mu_1, \mu_2)$,

$$\kappa_r(\mu_1, \mu_2) = \inf_{\mu} \int\limits_{[0,1]\times[0,1]} r(x,y)\, \mu(dx, dy),$$

where μ runs over all probability measures on $[0,1]\times[0,1]$ with marginal measures μ_1 and μ_2. We investigate the asymptotic behavior of the value $\kappa_r(\mathcal{P}_n^{Z^*}, P^{Z^*})$ and apply the obtained results to some problems of non-parametric goodness of fit.

References

1. Rüschendorf L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.* **4**, 912–923.

24

First-hitting-time Models and Threshold Regression

Mei-Ling Ting Lee

Biostatistics Division, School of Public Health
The Ohio State University

Abstract The first-hitting time (FHT) model has proved to be useful as an alternative model for time-to-event and survival data. On the basis of the FHT model, we introduce the threshold regression (TR) methodology. The threshold regression model has an underlying latent stochastic process representing a subject's latent health state. This health status process fluctuates randomly over time until its level reaches a critical threshold, thus defining the outcome of interest. The time to reach the primary endpoint or failure (death, disease onset, etc.) is the time when the latent health status process first crosses a failure threshold level. The effectiveness of threshold regression lies in how initial health status, hazards and the progression of disease are modeled, while taking account of covariates and competing outcomes. The threshold regression model does not require the proportional hazards assumption and hence offers a rich potential for applications.

25

Non-parametric Confidence Intervals for the Performability Function of Semi-Markov Systems

Nikolaos Limnios and Brahim Ouhbi

Université de Technologie de Compiègne, Compiègne, France &
Ecole Nationale Supérieure d'Arts et Métiers, Meknès, Maroc

Abstract: In this paper the evolution of the semi-Markov process with finite state space is considered. We develop a procedure for constructing confidence non-parametric intervals for the performability function. This investigation is based on the semi-Markov kernel estimator given in Ouhbi and Limnios [3]. This generalize the methods known for renewal function and point availability in renewal systems as well as for availabilty and reliabilty studies in semi-Markov systems.

Keywords and phrases: Non-paramertic confidence interval, Performability, Semi-Markov processes

25.1 Introduction

Performability of semi-Markov systems with finite state space and random holding time in each state are often used in many applied fields : reliability [1], economical studies [11], risk studies [10]. Construction of non-parametric confidence intervals for the point availability and reliability of semi-Markov systems was considered in Ouhbi and Limnios [6], it generalize the results obtained in Frees [8], Schneider [9] and Baxter et al.[7] for the renewal function and the point availability. In this paper, we propose a non-parametric method for constructing the confidence interval for the performability function of the semi-Markov systems. Our methodology is based on the results obtained in Limnios and Ouhbi [3].

25.2 Preliminaries

Let us consider a Markov renewal process (MRP) $(J,S)=(J_n,S_n)_{n\geq 0}$ defined on a probability complete space, where J_n is a Markov chain with values in E= $\{1,...,s\}$ which is the state space of the process and S_n are jumps times with values in R_+. The random variables J_0,J_1, ...,J_n,... are the consecutive states to be visited by the MRP and $X_1,X_2,...$ defined by $X_0=0$ and $X_n=S_n-S_{n-1}$, for $n\geq 1$, are the sojourn times in these states taking values in $[0,\infty)$.
A MRP can be completely determined if, we know its initial law and its transition probabilities defined respectively by :$P(J_0=k)=p(k)$ and

$$P[J_{n+1}=k,X_{n+1}\leq x|J_0,J_1,...,J_n,X_1,X_2,...,X_n]=Q_{J_nk}(x) \qquad (a.s.)$$

for all $x\in[0,+\infty)$ and $1\leq k\leq s$.
The probabilities $p_{ij}=Q_{ij}(\infty)$ $(=\lim_{t\to+\infty}Q_{ij}(t))$ are the transition probabilities of the embedded Markov chain J_n.
Let us, also consider the distribution function associated to sojourn time in state i before going to state j defined by:

$$F_{ij}(.)=\begin{cases} p_{ij}^{-1}\times Q_{ij}(.) & \text{if } p_{ij}>0 \\ 0 & \text{otherwise.}\end{cases}$$

So, we have:

$$P[J_n=j/J_0,J_1,....J_{n-1}=i]=p_{ij} \quad \text{for all} \quad n>0.$$

$$P[X_n\leq x/J_0,.....J_{n-1}=i,J_n=j]=F_{ij}(x) \text{ for all } n\geq 0 \text{ and } x\geq 0.$$

$$P[X_{n_1}\leq x_1,X_{n_2}\leq x_2,.....X_{n_k}\leq x_k/J_n,n\geq 0]=\prod_{i=1}^{k}F_{J_{n_i-1}J_{n_i}}(x_i) \qquad (a.s.)$$

for $0\leq n_1\leq n_2\leq ...\leq n_k$ and $x_i\geq 0$ for $i=1,...,k$.

The Markov renewal matrix, $\psi(t)$ is defined by

$$\psi(t)=\mathbb{E}[N(t)]=\sum_{l=0}^{\infty}Q^{(l)}(t),$$

where $N(t)$ is the counting process of transitions of the process up to time t and $Q_{ij}^{(1)}(t)=Q_{ij}(t)$ and for $l>1$, $Q^{(l)}(t)$ is the l^{th} convolution of Q(t) in the Stieltjes sense and

$$Q_{ij}^{(0)}(.)=\begin{cases} 1_{\{i=j\}}(t) & \text{if } t>0 \\ 0 & \text{otherwise.}\end{cases}$$

Let us recall that since the state space, E, is finite, $\psi(t)$ is element wise finite for every $t \geq 0$, i.e. the MRP is normal see [4]

The semi-Markov transition matrix function of the semi-Markov process, $(Z_t)_{t\geq 0}$, is defined by:

$$P_{ij}(t) = P[Z_t = j | Z_0 = i] = P[J_{N_t} = j | J_0 = i] \qquad i,\ j \in E,$$

It is known, cf. Pyke [13], that

$$P_{ij}(t) = 1_{\{i=j\}}(1 - \sum_{k=1}^{s} Q_{ik}(t)) + \sum_{k\in E} \int_0^t P_{kj}(t-s) Q_{ik}(ds).$$

By solving the above Markov renewal equation, cf. Limnios [1], it is seen that, in matrix notation,

$$P(t) = (I - Q(t))^{(-1)} * (I - diag(Q(t)\mathbf{1})), \tag{25.2.1}$$

where $diag(\cdot)$ is a diagonal matrix of i^{th} entry $\sum_{j=1}^{s} Q_{ij}(t)$ and $\mathbf{1} = (1, 1, ..., 1)^t$. The semi-Markov transition matrix is a very useful matrix function in studying the semi-markov processes asymptotic properties and in their applications see Ouhbi and Limnios [6].
Let $(Z_t)_{t\in\mathbb{R}_+}$ be a homogeneous semi-Markov process with finite state space $E = \{1, ..., s\}$,

$$W_t = \int_0^t L(Z_u) du,$$

$t \in \mathbb{R}_+$ where L is a measurable function taken to be a function of the occupied state and the holding time in that state so in the sequel

$$L(Z_u) = \sum_{n=0}^{\infty} g(J_n) 1_{\{S_n \leq u < S_{n+1}\}}$$

so

$$W(t) = \sum_{i\in E} \sum_{n=0}^{\infty} \int_0^t g(i) 1_{\{J_n = i, S_n \leq u < S_{n+1}\}} du.$$

Let us define the mean performance at time $t > 0$, denoted by $\overline{\Phi}(t) := EW(t)$. Then

$$\begin{aligned} \overline{\Phi}(t) := \mathbb{E}[W(t)] &= \sum_{i\in E} \int_0^t g(i) P[Z_u = i] du. \\ &= \sum_{i\in E} g(i) \int_0^t P_u(i) du. \end{aligned}$$

The problem here is to construct confidence intervals for $\overline{\Phi}(t)$ from a sample path truncated to the time interval $[0,T]$ of this process.
Let $(Z_t, 0 < t \leq T)$ be a given observation of the semi-Markov process on the fixed interval $[0,T]$. Let us define the following empirical estimator $\widehat{Q}_{ij}(t,T)$ for the semi-Markov kernel $Q_{ij}(t)$ at fixed time t

$$\widehat{Q}_{ij}(t,T) := \frac{1}{N_i(t)} \sum_{k=1}^{N(t)} \mathbf{1}_{\{J_{k-1}=i, J_k=j, X_k \leq t\}}. \tag{25.2.2}$$

Define also $\widehat{P}_{ij}(s,T)$, the estimator of $P_{ij}(t)$, by

$$\widehat{P}_{ij}(s,T) = (I - \widehat{Q}(s,T))^{(-1)} * (I - \widehat{H}(s,T))(i,j). \tag{25.2.3}$$

For the above estimators (25.2.2) and (25.2.3), the reader can find detailed results concerning asymptotic properties such as consistency and normality in [3, 6, 5].

Define now the following estimator for the mean performability $\overline{\Phi}_i(t) := \mathbb{E}_i \Phi(t)$,

$$\widehat{\overline{\Phi}}_i(t,T) := \sum_{j \in E} g(j) \int_0^t \widehat{P}_{ij}(s,T) ds. \tag{25.2.4}$$

In this section we give the asymptotic properties of this estimator when the time of observation tends to infinity for a unique trajectory .

Theorem 25.2.1
The estimator, $\widehat{\overline{\Phi}}_i(t,T)$ is uniformly strongly consistent in the sense that

$$\sup_{t \in [0,L]} |\widehat{\overline{\Phi}}_i(t,T) - \overline{\Phi}_i(t,T)| \longrightarrow 0 \quad a.s., \quad as \quad T \to \infty.$$

Theorem 25.2.2
For any fixed t, $t \in [0,\infty)$, $\sqrt{T}(\widehat{\overline{\Phi}}_i(t,T) - \overline{\Phi}_i(t,T))$ converges in law to a mean zero normal random variable with variance

$$\begin{aligned}
\sigma^2(t) &= \sum_{i=1}^{s} \sum_{j=1}^{s} \mu_{ii} \{ (W_{ij})^2 * Q_{ij} - (W_{ij} * Q_{ij})^2 \\
&+ \int_0^\infty [\int_0^\infty g(j)(x \wedge (t-u)) dA_i(u)]^2 dQ_{ij}(x) \\
&- [\int_0^\infty \int_0^\infty g(j)(x \wedge (t-u)) dA_i(u) dQ_{ij}(x)]^2 \\
&+ 2 \int_0^\infty W_{ij}(t-x) \int_0^\infty g(j)(x \wedge (t-u)) dA_i(u) dQ_{ij}(x) \\
&- 2(W_{ij} * Q_{ij})(t).(A_i * (g(j)(x \wedge .)))(t) \}
\end{aligned}$$

where for $t \in R_+$ *:*

$$A_i(t) = \sum_{k=1}^{s} \alpha_k g(i)\psi_{ki}(t), \quad \textit{and} \quad W_{kl}(t) = \sum_{i=1}^{s}\sum_{j\in U} \alpha_i(\psi_{ik} * \psi_{lj} * I_j)(t),$$

and $a \wedge b := \min\{a, b\}$.

25.3 Confidence intervals for the performability function

An estimator of $\sigma^2(t)$, denoted $\widehat{\sigma^2(t)}$, is obtained on replacing Q and ψ by $\hat{Q}$ and $\hat{\psi}$ respectively in the above expression. In the next theorem we show that $\hat{\sigma^2(t)}$ is strongly uniform consistent

Theorem 25.3.1
For any fixed t, $t \in [0, \infty)$*,The estimator,* $\widehat{\sigma^2(t)}$ *is uniformly strongly consistent in the sense that*

$$\sup_{t\in[0,L]} |\widehat{\sigma^2(t)} - \sigma^2(t)]| \longrightarrow 0 \quad a.s., \quad as \quad T \to \infty.$$

In summary, we have the following theorem

Theorem 25.3.2
$\frac{\sqrt{T}}{\widehat{\sigma(t)}}(\widehat{\overline{\Phi}}_i(t,T) - \overline{\Phi}_i(t,T))$ *converges in distribution to a standard normal random variate*

Hence for $\alpha \in (0, 1)$, an approximate $100(1 - \alpha)\%$ confidence interval for $E[W(t)]$ is

$$\widehat{\overline{\Phi}}_i(t,T) - z_{\frac{\alpha}{2}}\frac{\widehat{\sigma(t)}}{\sqrt{T}} \le \overline{\Phi}_i(t,T) \le \widehat{\overline{\Phi}}_i(t,T) + z_{\frac{\alpha}{2}}\frac{\widehat{\sigma(t)}}{\sqrt{T}}$$

where $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution.

References

1. N. Limnios and G. Oprişan (2001). *Semi-Markov Processes and Reliability*, Birkhäuser, Boston.
2. N. Limnios, B. Ouhbi and A. Sadek (2005), Empirical Estimator of Stationary Distribution For Semi-Markov Processes.*Com. Statist.: Theory and Methods*, **34**, 987-995.

3. B. Ouhbi and N. Limnios (1999). Non-parametric estimation for semi-Markov processes based on its hazard rate. *Statist. Infer. Stoch. Processes* **2(2)**, pp 151-173.
4. B. Ouhbi and N. Limnios (2001). The Rate of Occurrence of Failures for Semi-Markov Processes and Estimation. *Statist. Probab. Lett.,* **59(3)**, pp. 245-255.
5. Limnios, N., Ouhbi, B. (2003). Empirical estimators of reliability and related functions for semi-Markov systems. In *Mathematical and Statistical Methods in Reliability,* B. Lindqvist, K. Doksum (Eds.), World Scientific.
6. B. Ouhbi and N. Limnios (2003). Non-parametric Reliability Estimation of Semi-Markov Processes. *J. Statist. Plann. Infer.,* **109(1/2)**, pp. 155-165.
7. L.A. Baxter and L. Li (1994). Non-parametric Confidence Intervals for the Renewal Function and the Point Availability *Scan. J. of . Stat.,* **21**, pp. 277-287.
8. E.W. Frees(1986). Non-parametric Renewal function estimation. *Ann. Statist.,* **14**, pp. 1366-1378.
9. H. Schneider , B.S. Lin and C. O'Cinneide (1990). Comparison of non-parametric estimators for the renewal function. *Appl. Statist.,* **39**, pp. 55-61.
10. Ciardo, G., Marie, R.A., Sericola, B., Trivedi, K.S. (1990). Performability analysis using semi-Markov reward processes. *IEEE Trans. Comput.*, **C-39**, 1251–1264.
11. Papadopoulou, A. (2004). Economic rewards in non-homogeneous semi-Markov systems, *Commun. Statist. - Th. Meth.*, **33(3)**, pp 681–696.
12. R. Pyke and R. Schaufele (1966), The existence and uniqueness of stationary measures for Markov renewal processes. *Ann. Math. Statist.*, **37**: 1439-1462.
13. R. Pyke (1961), Markov renewal processes : definitions and preliminary properties. *Ann. Math. Statist.*, **32**: 1231-1241.
14. A. Scenski (1994), Cumulative operational time analysis of finite semi-Markov reliability models. *Reliability Engineering & system safety* , **44(1)**: 17-25.
15. S. Ross (1992), Applied probability models with optimization applications. *Dover New York.*
16. J. Janssen and N. Limnios(Eds) (1999), Semi-Markov models and Applications. *Kluwer Academic, Dordrecht.*

26

On Modeling and Estimability of Software Reliability

Tapan K. Nayak

Department of Statistics, George Washington University

Abstract: This talk will give an overview of statistical models and methods for software reliability estimation. General order statistics (GOS) models and non-homogeneous Poisson process (NHPP) models form two significant subclasses of the many software reliability models proposed in the literature. We shall discuss certain logical implications and criticisms of these two classes of models and estimability of the underlying parameters. The NHPP models with finite limit of the expected number of failures $m(\tau)$ as the testing time τ approaches ∞, have an important limitation. Specifically, the parameters of those models cannot be estimated consistently as the testing time approaches infinity. However, certain parameter based asymptotic properties of the maximum likelihood estimators can be obtained. Difficulties in estimating the unknown parameters from standard debugging data will be explained and an alternative experiment, called recapture debugging, which generates additional statistical information will be described.

26.1 Extended Abstract

Software has become a critical part of the operational technology of all major businesses, organizations, and government agencies. Releasing poor quality software is harmful to both the user and the reputation of the company developing the software. When a new software developed, it is tested with diverse inputs before its release, and whenever the software fails, efforts are made to identify the error (bug) that caused the failure and fix it. However, for large programs, detection and complete removal of all errors cannot be assured through software testing and debugging and the decision of when to release the software needs to be made based on its assessed reliability. Thus, statistical analysis

of software failure data and estimation of software reliability are important in software engineering.

Usually, the software is tested for a fixed amount of time τ with varied inputs, and the data consist of the observed failure times. Specifically, the random observables are: the number of failures during testing (R) and the successive failure times $0 \leq T_{(1)} \leq \cdots \leq T_{(R)} \leq \tau$. Typically, time is measured in processor running time, excluding time for debugging and idle time. Also, upon each failure efforts are made to detect and remove the error that caused the failure, so only the first occurrence of each error can be observed.

One class of models for analyzing software failure data, called the general order statistics (GOS) models (cf., Raftery, 1987), assume the following:

Assumption 1. Whenever the software fails, the error causing the failure is detected and corrected completely without inserting any new errors, i.e., the debugging process is perfect.

Assumption 2. The software initially contains an unknown number of errors, ν, and the detection times of those errors are independently and identically distributed (iid) with a common density $f_\theta(x)$, where $\theta > 0$ is an unknown parameter, possibly vector valued.

The GOS class includes the first software reliability model, introduced by Jelinski and Moranda (1972), where $f_\theta(x)$ is an exponential density function. Other GOS models may be viewed as modifications of the Jelinski-Moranda model.

The second class of models postulates that the failure counts follow a non-homogeneous Poisson process (NHPP). Let $M(t)$ denote the number of failures observed in the time interval $(0, t]$ and let $m(t) = E[M(t)]$. A NHPP model specifies the functional form of the intensity function $\lambda(t) = \frac{d}{dt}m(t)$, letting it depend on some unknown parameters. The Goel and Okumoto (1979) model is one of the earliest NHPP models for software reliability, in which the intensity function $\lambda(t)$ is assumed to be

$$\lambda(t) = \mu\theta \exp(-\theta t), \quad \mu > 0,\ \theta > 0.$$

Kuo and Yang (1996) classified NHPP models into two groups using $\lim_{t\to\infty} m(t)$. If the limit is finite, it is called NHPP-I, otherwise it is called NHPP-II. Moreover, NHPP-I processes can be expressed as mixtures of GOS processes. Specifically, if in the GOS model, ν is assumed to be a Poisson random variable with mean μ, it can be seen that $M(t)$ is NHPP-I with rate function $\lambda(t) = \mu f_\theta(t)$ and mean function $m(t) = \mu F_\theta(t)$ (cf. Langberg and Singpurwalla (1985), Musa, Iannino and Okumoto (1987), p. 269, or Kuo and Yang (1996)). Conversely, any given NHPP-I process with rate function $\lambda(t)$ and mean function $m(t)$ can be expressed as Poisson mixture of GOS processes with $\mu = \lim_{t\to\infty} m(t) < \infty$ and $f(t) = \lambda(t)\mu$. Examples of NHPP-I models include the models proposed by Goel and Okumoto (1979) where $\lambda(t) = \mu\theta e^{-\theta t}$, Yamada et al. (1983)

where $\lambda(t) = \mu\theta^2 te^{-\theta t}$, Goel (1985) where $\lambda(t) = \mu\theta\alpha t^{\alpha-1}\exp(-\theta t^\alpha)$, Littlewood (1984) where $\lambda(t) = \mu[1 - (\frac{\theta}{\theta+t})]^\alpha$, and Yamada and Osaki (1985) where $\lambda(t) = \mu/[1 + \alpha\exp(-\theta t)]$.

The NHPP models implicitly assume imperfect debugging as the number of failures in $[0, \infty)$ is unbounded for all values of the parameters μ and θ. The number of errors (initial or remaining) is not of direct interest in NHPP models. Moreover, all reliability measures depend only on the unknown parameters μ and θ (and not the data). Logically, software system reliability changes only when changes are made in the code, e.g., by detecting and fixing bugs. In between such changes, the reliability does not change (cf., Xie, 1991, p. 111). Thus, the reliability of a software changes only in jumps at times of debugging, and NHPP-I models are not consistent with that. Also, NHPP-I models assume independence of the failure process in non-overlapping intervals, but one would expect debugging activities at any time to affect the failure process subsequent to that time. We can show that no estimator of μ (or θ) converges in probability to the true value as the testing time τ converges to ∞, i.e., the parameters cannot be estimated consistently, which is a significant limitation of NHPP-I models.

For a GOS model, the assumption of perfect debugging is unrealistic. Also, the likelihood function can be very unstable and the MLE of ν can be infinite with positive probability. The nature of difficulties in estimating the parameters of a GOS model are exemplified by the Jelinski-Moranda model, the simplest GOS model. There, if one of the two parameters (ν or θ) is known, estimation of the other parameter is easy. For example, if θ is known, $\hat{\nu} = R/F_\theta(\tau)$ is the minimum variance unbiased estimator (MVUE) of ν and it is finite. However, estimating both ν and θ is quite difficult. This suggests that the information about ν and θ in the data is confounded. Letting $Y_i = T_{(i)} - T_{(i-1)}, i \geq 0, T_{(0)} = 0$, we note that $E(R)$ and $1/[E(Y_i)] = (\nu - i + 1)\theta, i = 1, \cdots, R$ are increasing functions of both ν and θ. So, ν and θ have similar effects on the random observables, and it is difficult to tell from the observed data whether ν is large and θ is small, θ is large and ν is small, or both are moderate.

It appears that for accurate inferences, additional information on one of the two parameters is necessary. Extending the GOS model, if the errors are assumed to be accessed according to iid renewal processes with renewal density $f_\theta(x)$, then extra information on θ would be obtained if the errors are not removed during testing and their repeat occurrence times are observed. This motivated Nayak (1988) to propose recapture debugging for estimating software reliability. Recapture debugging data can be modeled as a marked Poisson process and inferences about ν, θ, and various reliability measures can be made following standard statistical theory (cf., Nayak, 1988, 1991). In particular, the minimum variance unbiased estimators of many parametric functions can be obtained under suitable stopping rules. Statistical research on software re-

liability analysis has focused mainly on modeling and analyzing the standard debugging data. We believe the data collection aspect should not be ignored and other debugging methods for generating more useful information should be explored.

References

1. Goel, A. L. (1985). Software reliability models: assumptions, limitations, and applicability, *IEEE Trans. Software Eng.*, SE-11, 1411-1423.

2. Goel, A. L. and Okumoto, K. (1979). Time-dependent error detection rate model for software reliability and other performance measures, *IEEE Trans. Reliability*, R-28, 206-211.

3. Jelinski, Z. and Moranda, P.M. (1972). Software reliability research, In *Statistical Computer Performance Evaluation*, (ed. W. Freiberger), pp. 465-484, Academic Press, New York.

4. Kuo, L. and Yang, T. Y. (1996). Bayesian computation for nonhomogeneous Poisson processes in software reliability, *J. Amer. Statist. Assoc.*, 91, 763-773.

5. Langberg, N. and Singpurwalla, N. D. (1985). A unification of some software reliability models, *SIAM J. Sci. & Statist. Computing*, 6, 781-790.

6. Littlewood, B. (1984). Rationale for a modified Duane model, *IEEE Trans. Reliability*, R-33, 157-159.

7. Musa, J. D., Iannino, A. and Okumoto, K. (1987). *Software Reliability: Measurement, Prediction, Application,* McGraw-Hill, New York.

8. Nayak, T. K. (1988). Estimating population size by recapture debugging. *Biometrika* 75, 113-120.

9. Nayak, T. K. (1991). Estimating the number of component processes of a superimposed process. *Biometrika* 78, 75-81.

10. Raftery A. E. (1987). Inference and prediction for a general order statistic model with unknown population size, *J. Amer. Statist. Assoc.*, 82, 1163-1168.

11. Xie, M. (1991). *Software Reliability Modelling*, World Scientific Publishing, Singapore.

12. Yamada, S. and Osaki, S. (1985). Software reliability growth modeling: models and applications, *IEEE Trans. Software Eng.*, SE-11, 1431-1437.

13. Yamada, S., Ohba, M. and Osaki, S. (1983). S-shaped reliability growth modeling for software error detection, *IEEE Trans. Reliability*, R-32, 475-478.

27

On Measures of Information and Divergence: Some Recent Developments

Takis Papaioannou

University of Piraeus, Greece

27.1 Introduction[1]

Measures of information appear everywhere in probability and statistics. They also play a fundamental role in communication theory. They have a long history since the papers of Fisher, Shannon, and Kullback. There are many measures each claiming to capture the concept of information or simply being measures of (directed) divergence or distance between two probability distributions. Also there exist many generalizations of these measures. One may mention here the papers of Lindley and Jaynes who introduced entropy based Bayesian information and the maximum entropy principle for determining probability models, respectively.

Broadly speaking there are three classes of measures of information and divergence: Fisher-type, divergence-type, and entropy (discrete and differential)-type measures. Some of them have been developed axiomatically (see, for example, Shannon's entropy and its generalizations), but most of them have been established operationally in the sense that they have been introduced on the basis of their properties.

There have been several phases in the history of information theory: Initially we have (i) the development of generalizations of measures of information and divergence (f−divergence, (h−f)−divergence, hypo-entropy, etc), (ii) the synthesis (collection) of properties they ought to satisfy, and (iii) attempts to unify them. All this work refers to populations and distributions. Later on

[1]Part of this work was done while the author was Visiting Professor at the University of Cyprus.

we have the emergence of information or divergence statistics based on data or samples and their use in statistical inference primarily in minimum "distance" estimation and for the development of asymptotic tests of goodness of fit or model selection criteria. Lately we have a resurgence of interest on measures of information and divergence and they are used in many places, in several contexts and in new sampling situations. Some of these ideas are discussed below.

The measures of information and divergence enjoy several properties (non-negativity, maximal information, sufficiency etc) and statisticians do not agree on all of them. There is a body of knowledge known as statistical information theory which has made many advances but not achieved a wide acceptance and application. The approach is more operational rather than axiomatic as it is the case with Shannon's entropy.

There are several review papers which discuss the above points. We mention the following: Kendall (1973), Csiszar (1977), Kapur (1984), Aczel (1986), Papaioannou (1985 and 2001), Soofi (1994 and 2000). The aim of this talk is to develop a general appreciation on the meaning and uses of various properties rather than on their mathematical content.

We shall present some recent developments on measures of information and divergence as follows:

1. We shall discuss the properties of sub-additivity

$$I(X,Y) \leq I(X) + I(Y),$$

super-additivity

$$I(X,Y) \geq I(X) + I(Y),$$

and conditional inequality

$$I(X|Y) \leq I(X),$$

where I is any measure of information or divergence. We shall expand on their implications to statistical theory (measures of correlation or dependence). For recent papers see Papaioannou and Ferentinos (2005) and Micheas and Zografos (2006).

2. We shall present several inequalities involving measures of information and divergence.

3. We shall review information and divergence under censoring both informative and noninformative. We shall discuss the 'acid test" property imposed by the censoring process as well as various interrelationships between measures of information or divergence in several censoring situations. One recent use of

informativity has to do with Fisher information in weighted distributions, in order statistics, and in record statistics (Arnold et al. (1998), Abo-Eleneen and Nagaraja (2002), Park and Zheng (2004)). A weighted distribution is a distribution which has a density in its pdf, i.e., if $f(x,\theta)$ is a density, we use a density proportional to $w(x)f(x,\theta)$ to make inferences about θ. Weighted distributions are used to model ascertainment bias. Iyengar et al. (1999) studied conditions under which the Fisher information about θ obtained from a weighted distribution, is greater than the same information obtained from the original density $f(x,\theta)$, where $f(x,\theta)$ belongs to an exponential family. This is clearly a result on information. Thus, there are cases where *the Fisher information about θ contained in an order statistic, is greater than the same information contained in a single observation.* This follows from the fact that the distribution of an order statistic is a weighted distribution. It turns out that for the normal distribution with σ^2 known, $I^F_{X_{(k)}}(\mu) \geq I^F_X(\mu)$, where $X_(k)$ is the k^{th} order statistic of a random sample from $X_1, X_2, ..., X_k$ from $N(\mu, \sigma^2)$. This result is in agreement with our intuition, since the order statistic essentially involves the whole sample. Other interesting informativity applications appear with the *residual lifetime* of a stationary renewal process or with truncated distributions. For details see Iyengar et al. (1999). Several studies have shown that the tails of an ordered sample from a symmetric distribution contain more Fisher information about the scale parameter than the middle portion. Zheng and Gastwirth (2000, 2002) examined the Fisher information about the scale parameter in two symmetric fractions of order statistics data from four symmetric distributions. They showed that for the Laplace, logistic and normal distribution, the extreme tails usually contain most of the Fisher information about the scale parameter, while the middle portion is less informative. For the Cauchy distribution the most informative two symmetric fractions are centered at the 25^{th} and 75^{th} percentile

4. Similar results as in the previous paragraph exist when we deal with truncated data, and in particular samples from truncated exponential distributions. Bayarri et al. (1989) give conditions under which for the Fisher information

$$I(X) < I(Y) \quad \text{or} \quad I(X) = I(Y) \quad \text{or} \quad I(X) > I(Y),$$

where X follows an arbitrary exponential distribution of the form

$$f(x,\theta) = a(x)exp(b(\theta)u(x)/c(\theta), \theta \in \Theta$$

and Y follows the truncated distribution

$$g(y,\theta) = \begin{cases} f(y,\theta)/s(\theta), & \text{for} \quad y \in S \\ 0, & \text{otherwise} \end{cases}$$

The set S, a subset of the sample space of X, is the truncation or selection set with $s(\theta) = P_\theta(X \in S) = \int_S f(x,\theta)dx$. A selection sample from the right tail of the normal distribution contains less Fisher information about the mean than an unrestricted random sample when the variance is known, but more information about the variance than an unrestricted random sample when the mean is known.

5. We shall look at recent work involving evaluations of measures of information. There is a recent upsurge of this kind of evaluations starting with the work of Guerrero and continuing with the work of Zografos, Nadaraya, and Cavanaugh and Shumway.

6. We shall also discuss model comparisons and model selection criteria using measures of information and divergence. The work here originates with the Akaike information criterion (AIC).

7. Time permitting we shall discuss information in frailty models.

References

1. Abo-Eleneen, Z. and Nagaraja, H. (2002). Fisher information in an order statistic and its concomitant. Annals of the Institute of Statistical Mathematics, 54, 667-680.

2. Aczel, J. (1986). Characterizing information measures: Approaching the end of an era. In Uncertainty in Knowledge-Based Systems. Lecture Notes in Computer Science. (Eds., B. Bouchon and R.R. Yager), pp. 359-384. Spinger-Verlag, New York.

3. Arnold, B., Balakrishnan, N. and Nagaraja, H. (1998). Records. Wiley, New York.

4. Bayarri, M. J., DeGroot, M. H. and Goel, P. K. (1989). Truncation, information and the coefficient of variation. In Contributions to Probability and Statistics. Essays in Honor of Ingram Olkin (Gleser, L., Perlman, M., Press, S. J. and Sampson, A., Eds.), Springer Verlag, New York, 412-428.

5. Cavanaugh, J.E. and Shumway, R.H. (1996). On computing the expected Fisher information matrix for state-space model parameters. Statistics and Probability Letters, 26, 347-355.

6. Csiszar, I. (1977). Information measures: A critical review. Transactions of the 7th Prague Conference on Information, Theory of Statistical Decision Functions, and Random Processes, pp 73-86, Prague , 1974, Academia.

7. Guerrero-Consumano, J.L. (1996). A measure of total variability for the multivariate t distribution with applications to finance. Information Sciences, 92, 47-63.

8. Iyengar, S., Kvam, P. and Singh, H. (1999). Fisher information in weighted distributions. The Canadian Journal of Statistics, 27, 833-841.

9. Kapur,J.N. (1984). A comparative assessment of various measures of divergence. Advances in Management Studies, 3, 1-16.

10. Kendall, M.G. (1973).Entropy, probability and information. International Statistical Review, 11, 59-68.

11. Micheas,A.C. and Zografos,K (2006). Measuring stochastic dependence using -divergence. J. of Multivariate Analysis, 97, 765-784.

12. Nadarajah,S. (2006). Information matrices for Laplace and Pareto mixtures. Computational Statistics and Data Analysis. 50, 950-966.

13. Papaioannou, T. (1985). Measures of information. Encyclopedia of Statistical Sciences, Kotz and Johnson, Eds., 5, Wiley, New York, 391-397.

14. Papaioannou, T. (2001). On distances and measures of information: a case of diversity. Probability and Statistical Models with applications, C.A.Charalambides, M.V. Koutras and N. Balakrishnan,Eds., Chapman and Hall/CRC, London, 503-515.

15. Papaioannou, T. and Ferentinos, K. (2005). On two forms of Fisher's measure of information. Communications in Statistics - Theory and Methods, 34, 1461-1470.

16. Soofi, E.S. (1994). Capturing the intangible concept of information. J. of the American Statistical Association, 89, 1243-1254.

17. Soofi, E.S. (2000).Principal information theoretic approaches. J. of the American Statistical Association, 95, 1349-1353.

18. Park, S. and Zheng, G. (2004). Equal Fisher information in order statistics. Sankhya; The Indian Statistical Journal, Series B, 66, 20-34.

19. Zheng, G. and Gastwirth, J. L. (2000). Where is the Fisher information in an ordered sample? Statistica Sinica, 10, 1267-1280.

20. Zheng, G. and Gastwirth, J. L. (2002). Do tails of symmetric distributions contain more Fisher information about the scale parameter? Sankhya: The Indian Journal of Statistics, Series B, 64, 289-300.

21. Zografos, K. and Nadarajah, S. (2005). Expressions for Renyi and Shannon entropies for multivatiate distributions. Statistics and Probability Letters, 71, 71-84.

28

A Model Free Approach to Combining Diagnostic Markers

Ruth Pfeiffer and Efstathia Bura

National Cancer Institute and George Washington University, USA

Abstract: A popular summary measure of the discriminatory ability of a single continuous diagnostic marker for binary disease outcomes is the receiver-operator characteristics curve (ROC). For most diseases however, single biomarkers do not have adequate sensitivity or specificity for practical purposes. We present an approach to combine several markers into a composite diagnostic test without assuming a model for the distribution of the predictors. Using sufficient dimension reduction techniques, we replace the predictor vector with a lower-dimensional version, obtained through linear transformations of biomarkers, without loss of information. We show how to combine the linear transformations using their asymptotic properties into a scalar diagnostic score whose performance can be assessed by the ROC curve. In the special case that a single linear combination of the markers contains sufficient information for the outcome, this approach results in the same marker combination obtained by Su & Liu (1993) that maximises the area under the ROC curve. The asymptotic distribution of the left singular vectors of a consistent estimate of an asymptotically normally distributed random matrix is derived, which provides the basis for an asymptotic chi-squared test to assess individual biomarker contribution to the diagnostic score.

Keywords and phrases: Dimension reduction; Likelihood ratio; NHANES III; Random matrix; SAVE; SIR; Singular value decomposition.

28.1 Introduction

The emerging field of clinical proteomics involves the discovery and identification of new biomarkers that may aid in diagnosis of disease, prediction of clinical outcome, and therapeutic efficacy for a host of diseases, including cancer, cardiovascular and mental conditions. In an ideal setting one would obtain a single

marker with very high specificity and sensitivity, or predictive ability, for the desired outcome. However, such high performance markers are yet to be found for many diseases, including many cancers. A more realistic approach is to combine several markers of modest individual discriminatory ability to improve discrimination and performance for diagnosis or screening applications.

A popular summary measure of the discriminatory ability of a single diagnostic marker in the setting of a binary disease outcome is the receiver-operator characteristics curve (ROC) that plots sensitivity against (1-specificity) (true vs. false positivity) for all thresholds that could have been used to define "test positive." Two tests can be compared by calculating the difference between the areas under their two ROC curves (AUCs), with the larger area corresponding to the "better" test. In order to use the ROC curve as a measure of diagnostic accuracy for several markers, a scalar function of the marker values is needed. If one knows the joint distribution of all the markers in a panel in both the case and the control populations, the likelihood ratio (LR) statistic is such a scalar and provides the most powerful means of combining the markers (McIntosh & Pepe, 2002). However, the LR approach may be sensitive to violations of the distributional assumptions and does not allow one to easily assess the contributions of individual markers. Assuming multivariate normality of the markers, several authors have derived a diagnostic score consisting of a single linear combination of the marker values that maximises the area under the ROC curve (Su & Liu, 1993), or the sensitivity over a range of specificities (Liu et al., 2005). Pepe & Thompson (2000) relax the assumption of multivariate normality and find a linear marker combination that maximises a distribution free estimate of the AUC. While their approach is attractive, since it can be adapted in order to maximise the partial area under the curve (McGlish, 1989) and to incorporate covariates, it is computationally difficult when one wishes to combine more than two markers.

In this paper, we extend the approach of Su & Liu (1993) by relaxing the assumption of multivariate normality of the markers, and, more importantly, by identifying a sufficient number of linear combinations of markers that can be combined into a diagnostic score, using two sufficient dimension reduction methods, Sliced Inverse Regression (SIR) and Sliced Average Variance Estimation (SAVE). First we provide general background on SAVE and SIR. The SAVE procedure, applied to discrimination between two populations, defines a sufficient subspace for discrimination. We combine projections of the original markers into this subspace that are uncorrelated by construction and asymptotically normally distributed via a likelihood ratio statistics to obtain a scalar diagnostic score for discrimination. Neither SAVE nor SIR require the specification of a model for the relationship between the markers and the outcome. We also show that SIR, a first order moment method, results in the same linear combination that Su and Liu (1993) obtained. Even though aggregate predictors

such as linear combinations effectively combine information from all markers, individual marker contribution to a diagnostic panel may be of significant interest. We also derive the asymptotic distribution of the left singular vectors of consistent estimates of an asymptotically normally distributed random matrix. This general result provides the basis for an asymptotic chi-squared test to assess which of the original markers do not contribute to the SAVE predictors and thus are superfluous for computing the diagnostic score. This test can be applied to assess variable contribution to any linear combination of predictors whose coefficients are the elements of the left singular vectors of an asymptotically normal random matrix. Hence, it is not limited to SAVE but can be used in all dimension reduction methods that are based on the estimation of a kernel matrix that is asymptotically normal. We use simulations to assess the performance and robustness of the proposed scalar discrimination scores based on SAVE and SIR and to test for marker contributions to the diagnostic score. A data example is presented in and we conclude with a discussion of our results.

28.2 Dimension Reduction Methods based on Inverse Regression

We denote the marker values by $X = (X_1, \ldots, X_p)^T$ and the outcome variable by Y. Before we focus on the setting of binary outcomes $Y = 0$ for nondiseased and $Y = 1$ for diseased individuals, we provide some general framework for inverse regression that applies to both continuous and discrete Y. A data reduction formulation that accounts for the correlation among markers, is to assume there exists a $p \times d$, $d \leq p$, matrix η so that the $p \times 1$ predictor vector X can be replaced by the $d \times 1$ predictor vector $\eta^T X$ without loss of information for the regression of Y on X. Most importantly, if $d < p$, then sufficient reduction in the dimension of the regression is achieved. The linear subspace $S(\eta)$ spanned by the columns of η is a dimension-reduction subspace (Li, 1991, 1992) and its dimension denotes the number of linear combinations of the components of X needed to model Y.

The central dimension-reduction subspace, denoted by $S_{Y|X}$ (Cook, 1996, 1998) is the intersection of all dimension-reduction subspaces for $F(Y|X)$ and is trivially the smallest dimension-reduction subspace when it exists. The dimension $d = \dim(S_{Y|X})$ is called the structural dimension of the regression of Y on X and can take on any value in the set $\{0, 1, \ldots, p\}$.

The estimation of the central subspace is based on finding a kernel matrix Ω_x so that $S(\Omega_x) \subset S_{Y|X}$. This can be done by first moment methods such as SIR (Li, 1991) with $\Omega_x = \mathrm{cov}(\mathrm{E}(X|Y))$, or second moment methods including SAVE (Cook & Weisberg, 1991), with $\Omega_x = \mathrm{E}(\mathrm{cov}(X) - \mathrm{cov}(X|Y))^2$. SAVE is the most comprehensive dimension reduction method as it gains information from both the inverse mean function and the differences of the inverse covariances.

28.2.1 SIR and SAVE for binary outcomes Y

Let $\mu_x = \mathrm{E}(X)$ and $\Sigma_x = \mathrm{var}(X)$. Also let the conditional moments on the binary disease status be $\mu_{x|j} = \mathrm{E}(X|Y = j)$ and $\Sigma_{x|j} = \mathrm{var}(X|Y = j)$ for $j = 0, 1$. We assume that Σ_x and $\Sigma_{x|j}$, $j = 0, 1$ are nonsingular. Denote the standardised predictors by $Z = \Sigma_x^{-1/2}(X - \mathrm{E}(X))$, and the conditional means and variances, for the binary response Y, by $\mu_j = \mathrm{E}(Z|Y = j)$ and $\Sigma_j = \mathrm{var}(Z|Y = j)$, $j = 0, 1$.

The subspace spanned by the columns of the differences in the first two moments of $Z|Y$, $\nu = \mu_1 - \mu_0$ and $\Delta = \Sigma_1 - \Sigma_0$, is $S(\Delta, \nu)$. Cook & Lee (1999) showed that the SAVE kernel matrix based on the standardised predictors is $\Omega_{SAVE} = \Omega_z = (\Delta, \nu)$, and hence that $S_{SAVE} = S(\Delta, \nu)$ contains some or all the sufficient linear combinations that can replace the predictor vector Z in the regression of Y on Z, under the two moment conditions stated in the previous section. Furthermore, when the conditional distribution of $Z|Y$ is normal, then $S_{SAVE} = S_{Y|Z}$. They also showed that $\Omega_{SIR} = \nu$, that is, $S_{SIR} = S(\nu) \subset S_{SAVE}$.

In implementing SAVE or SIR, ν and Δ are replaced by the corresponding sample moments, $\hat{\nu} = \hat{\Sigma}_x^{-1/2}(\bar{x}_1 - \bar{x}_0)$ and $\hat{\Delta} = \hat{\Sigma}_x^{-1/2}(\hat{\Sigma}_{x|1} - \hat{\Sigma}_{x|0})\hat{\Sigma}_x^{-1/2}$ to yield $\hat{S}_{SAVE} = S(\hat{\Delta}, \hat{\nu})$, a $k \times (k + 1)$ matrix, and $\hat{S}_{SIR} = S(\hat{\nu})$, a $k \times 1$ vector. The latter has obviously dimension at most 1. To assess the dimension $d = \dim(\mathcal{S}_{\mathrm{SAVE}})$, a test statistic for dimension can be used that is a function of the singular values of the $\hat{\Omega}_z = \hat{\Omega}_{SAVE} = (\hat{\Delta}, \hat{\nu})$, or $\hat{\Omega}_{SIR} = \hat{\nu}$, depending on the method used. Cook & Lee (1999; Theorem 3) provide details on the test statistic for SAVE in the binary outcome setting and show that it has an asymptotic weighted chi-squared distribution. Approximate p-values can be obtained using a result, for example, by Wood (1989). The SIR test statistic for dimension has an asymptotic chi-squared distribution (Li, 1991; Bura & Cook, 2001). In both cases, the estimation is carried out by performing tests of $H_0 : p = d$ against $H_a : p > d$ sequentially, starting at $d = 0$, which corresponds to independence of Y and X, and adding unit increments to d until we cannot reject the null at a prespecified α level. For binary classification problems, SIR can estimate at most one basis element of $S_{Y|X}$.

28.3 Combining diagnostic markers using SIR and SAVE

As mentioned in the introduction, a popular measure to quantify performance of a diagnostic test is the ROC curve, that plots sensitivity of a test against 1-specificity for all possible thresholds that could be used to define "test positive." The most widely used summary measure for the ROC curve is the area under the curve (AUC), defined as $AUC = \int_0^1 ROC(t)dt$. The AUC can also be expressed

as the probability that the scalar diagnostic scores S_i for the cases (S_1) and controls (S_0) are correctly ordered, that is $AUC = \mathrm{pr}(S_1 > S_0)$. The SIR single linear combination of the markers can be used directly as a diagnostic score. However, for SAVE a scalar needs to be derived if the dimension is estimated to be larger than one.

28.3.1 Relation of SIR to existing results on linear combinations of multiple markers

In the case where the biomarkers are normally distributed for both controls and cases, SIR results in the same linear combination that was obtained by Su and Liu (1993) by maximising

$$AUC(a) = \mathrm{pr}(a'X_1 > a'X_0) = \Phi\Big\{\frac{a'(\mu_{x|1} - \mu_{x|0})}{(a'\Sigma_x a)^{1/2}}\Big\},$$

as a function of a. The maximiser is $a^* = c\Sigma_x^{-1}(\mu_{x|1} - \mu_{x|0})$, for any constant c, with $\Sigma_x = \Sigma_{x|0} + \Sigma_{x|1}$. As indicated in section 28.2.1, $S_{SIR} = S(\nu)$ where $\nu = \Sigma_x^{-1/2}(\mu_{x|0} - \mu_{x|1})$. The linear combination obtained from SIR, under only the linearity condition as defined in §2, is thus proportional to the linear combination obtained by Su and Liu, and consequently maximises the AUC. When $\Sigma_0 = \Sigma_1 = \Sigma$, this linear combination is proportional to the linear combination obtained by LDA, $\Sigma_x^{-1}(\mu_{x|1} - \mu_{x|0})X$.

28.3.2 Combining diagnostic markers using SAVE

McIntosh & Pepe (2002) showed that among all possible functions of X, the likelihood ratio function $LR = LR(X) = \mathrm{pr}(X|Y = 1)/\mathrm{pr}(X|Y = 0)$ has an ROC curve with the maximal AUC. Thus, if the joint distribution of the markers X is known among cases and controls, an optimal scalar marker score can be computed using the LR function. If the markers in the case and control populations are multivariate normal, that is, for $j = 0, 1$, $(X|Y = j) \sim MVN(\mu_{x|j}, \Sigma_{x|j})$, then $\log LR = \log\{f_1(x)/f_0(x)\} = C + 1/2x'(\Sigma_{x|0}^{-1} - \Sigma_{x|1}^{-1})x + x'(\Sigma_{x|1}^{-1}\mu_{x|1} - \Sigma_{x|0}^{-1}\mu_{x|0})x$. The LR statistic for multivariate normal predictors is thus fully characterized by $(\Sigma_{x|0}^{-1} - \Sigma_{x|1}^{-1})$ and $(\Sigma_{x|1}^{-1}\mu_{x|1} - \Sigma_{x|0}^{-1}\mu_{x|0})$, which also determine Ω_{SAVE}. The SAVE predictors completely capture the discriminatory information contained in the LR statistic.

Motivated by the above argument we propose the following approach for combining diagnostic test results based on the SAVE predictors:

1. We first estimate the dimension d of $\hat{S}_{\mathrm{SAVE}}$, and then, given d, the corresponding linear combinations, $x_1^* = \Sigma_x^{-1/2}U_{z1}X, \ldots, x_d^* = \Sigma_x^{-1/2}U_{zd}X$.

2. Diaconis & Freedman (1994) and Hall & Li (1993) showed that, under mild conditions, low-dimensional projections of high-dimensional data are approximately Gaussian. We thus assume that the estimated SAVE predictors, X^*, are approximately normally distributed with mean μ^* and variance-covariance structure Σ^*, regardless of the distribution of the original markers. In addition, the SAVE predictors are orthogonal, and thus independent by construction, with $\Sigma^* = diag(\sigma_i^*)$. We use the sample moments of X^* to estimate μ^* and Σ^*.

3. To compute a diagnostic score S for the SAVE predictors $x_1^*, \ldots, x_d^*$, we use the LR statistic (McIntosh & Pepe, 2002), based on the product of m independent univariate normal distributions, that is

$$S(x^*) = LR(x^*) = \frac{f_1(x^*)}{f_0(x^*)} = \frac{\prod_{j=1}^{d} \phi(x_j^*; \mu_{1j}^*, \sigma_{1j}^*)}{\prod_{j=1}^{d} \phi(x_j^*; \mu_{0j}^*, \sigma_{0j}^*)}$$

   where ϕ denotes the univariate normal density function.

The key advantage of using the SAVE predictors in the LR statistic instead of the markers directly is that except for the linearity and constant variance conditions, no other specific distributional assumption for the markers is needed.

29

Sequential Analysis Using Item Response Theory Models

Véronique Sébille

Laboratoire de Biostatistiques, Faculté de Pharmacie
Université de Nantes

29.1 Abstract

Many clinical trials attempt to measure health-related Quality of Life (QoL) that reflect patient's perception of his or her well-being and satisfaction with therapy. QoL endpoints are often used in non-comparative or comparative clinical trials which are commonly designed to evaluate therapeutic efficacy as well as further investigation of the side-effects and potential risks associated with therapy. Early stopping of such trials in case of beneficial or deleterious effect of the treatment on QoL is an important matter. The use of sequential tests taking into account the specific nature of QoL data seems to provide a powerful method to detect therapeutic effects [1].

QoL is usually evaluated using self-assessment questionnaires and responses to the items are usually combined into QoL scores assumed to be normally distributed. However, these QoL scores are rarely normally distributed and usually do not satisfy a number of basic measurement properties. An alternative is to use item response theory (IRT) models such as the Rasch model for binary items which takes into account the categorical nature of the items [2]. In this framework, the probability of response of a patient on an item depends upon different kinds of parameters: the "ability level" of the person (which reflects his/her current QoL) and a set of parameters characterizing each item.

Sequential analysis and mixed Rasch models assuming either known or unknown items parameters values were combined in the context of phase II, phase III comparative clinical trials. The statistical properties of two sequential tests [3], the Sequential Probability Ratio Test (SPRT) and the Triangular Test (TT) are compared using mixed Rasch models and the traditional method based on QoL scores by means of simulations. An example of the use of combining sequential analysis and IRT modelling methodologies is also presented on data

from a comparative phase III clinical trial in oncology.

References

1. Sébille V. and Mesbah M. (2006). Sequential Analysis of Quality of Life Rasch Measurements. In Nikulin M, Commenges D, Huber C. (eds.) *Probability, Statistics and Modelling in Public Health.* Springer - Statistics.

2. Fisher, G.H. and Molenaar, I.W. (1995). *Rasch Models, Foundations, Recent Developments, and Applications.* New-York: Springer-Verlag.

3. Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials.* Chichester: Wiley, revised 2nd ed.

30

The Utility of Reliability and Its Coherent Elicitation

Nozer D. Singpurwalla

The George Washington University
Washington, DC 20052, USA

Abstract: Our interest in utility theory has been sparked by its frequent mention in quality of life studies, and its total lack of mention in reliability theory and survival analysis. It seems to us that if the aim of reliability theory is to assist decision making in the context of system design, then the utility of reliability should play a key role. Yet little, if any, has been written on this topic. The aim of this paper is two-fold. The first is to articulate on the issue of the utility of reliability and the general nature of the utility function. The second is to describe an approach for the coherent assessment of utilities. Whereas the first aim is specific to reliability, the second is more general and should appeal to all decision theorists.

In order to formalize the role of the utility of reliability for decision making in the context of system design, we need to distinguish between reliability and survivability. The former is to be seen as a chance or a propensity, and the latter as one's uncertainty about propensity. This distinction is in keeping with the essential spirit of de Finetti's famous theorem on exchangeable Bernoulli sequences wherein he links the objective and the subjective interpretations of probability. Thus to us here, *reliability is a chance, not a probability.*

For the coherent elicitation of utilities we lean on the literature on quality of life studies wherein the Rasch Model of item response theory plays a prominent role. It has been claimed, but the specifics have not been given, that the Rasch Model can be used to assess utilities. In this talk, following a brief overview of the meaning of utility, we outline a statistical procedure for the coherent assessment of utilities using a binary response model, like the Rasch Model. Besides coherence, the virtue of using statistical approaches for utility assessment is that now we can provide measures of uncertainty about utility, and mechanisms for updating utilities. Both these possibilities have not, to the best of our knowledge, been covered in the literature on utility theory for

decision making.

This talk is expository and outlines work in progress.

31

Adaptive Designs for Group Sequential Clinical Survival Experiments

Eric V. Slud

Statistics Program, University of Maryland College Park, USA

Abstract: Randomized two-group clinical survival experiments now commonly allow at least one interim look, enabling possible early stopping to meet ethical concerns. Various authors have also studied the possibility of interim design modifications to adapt to unexpected accrual or control-group mortality rates. This paper formulates trial design as a decision theoretic problem with a large class of loss functions, in the setting of a statistic with the asymptotic behavior of Brownian motion with drift, as in Leifer and Slud (2002). A more general observation process arises in the case of adaptive designs allowing the option of continued followup without new accrual past an interim look, as was introduced in Koutsoukos, Rubinstein and Slud (2000). Some optimal two-look designs are displayed, and both types of adaptation are given a unified form.

Keywords and phrases: Accrual stopping; Backward induction; Bayesian decision theory; behavioral decision rule; Loss function; Stopping boundaries

31.1 Introduction

Group sequential designs are designs in which experimental data on two-group treatment comparisons can be scrutinized at a finite number of interim look-times with the possibility of early stopping of the experiment in such a way as to maintain a prescribed experimentwise significance level. Such designs first appeared for two-group randomized clinical trials with normally distributed quantitative responses in the mid-1970's. After a few years, methods appeared which took explicit account of the staggered entry, followup time, and delayed response of clinical trials with survival-time endpoints. By the early 1980's, such methods were firmly established theoretically. Work of Tsiatis (1982) showed that the repeatedly computed logrank-numerator statistic at a series

of fixed scheduled interim look-times would under standard conditions behave in large two-sample trials as a sequence of independent-increment Gaussian variables, with mean 0 under the null hypothesis $\mathbf{H}_0$ of no treatment effect and with steady positive drift proportional to variance under local proportional-hazard alternatives. Slud and Wei (1982) showed how variance increments could be progressively estimated and at the same time early stopping could be accommodated under an *α-spending schedule*. In a (one-sided) trial of sample size n, with the statistic $S_k/\sqrt{n}$ calculated at the k'th look-time t_k, a threshold or boundary b_k is used to stop the trial early with rejection of $\mathbf{H}_0$ if $S_k/\sqrt{n} \geq b_k$, where b_k is found inductively, in terms of the estimated large-sample variance V_k of $S_k/\sqrt{n}$, to satisfy

$$\alpha_k = Pr(S_j/\sqrt{n} < b_j \quad \text{for} \quad 1 \leq j < k, \quad S_k/\sqrt{n} \geq b_k) \tag{31.1.1}$$

and where the values $\alpha_1, \ldots, \alpha_K$ are prescribed and sum to the experimentwise significance level α. The times at which interim looks might be taken in this setup can be allowed to be random stopping-times, e.g., to be level-crossing times for the proportional-hazard parameter's *information*, which is proportional to the logrank variance and thus also to the number of observed failure events. Moreover, the choice of the specific value α_k need not be made until the $k-1^{\text{th}}$ look-time (Lan and DeMets 1983). The asymptotic theory underlying this extension was given by Slud (1984) and other authors, establishing that under local (contiguous) proportional-hazard alternatives the repeatedly computed logrank statistic considered as a stochastic process behaves asymptotically in large samples like a time-changed Brownian motion with drift. The history of these developments from the viewpoint of trial design, along with practical recommendations on the choice among early-stopping designs as of 1984, can be found in Fleming et al (1984).

Later research on the specification of early-stopping boundaries included generalizations beyond our scope here (more general statistics, adjustment for covariates, modified formulations of repeated significance testing, etc.), but also developed optimization methods: Tsiatis and co-authors restricted attention to parametrically restricted families of boundaries and computed the ones which minimized expected trial duration over boundaries with prescribed size and (average) power against specified alternative(s); while Jennison (1987) undertook a brute-force (grid-search) computation of optimal boundaries in the sense of minimizing a weighted linear combination of type-II error probabilities and expected sample sizes over specified alternatives, for given significance level.

Clinical investigators often find at the times of interim looks in clinical trials that planned accrual goals have not been met, and sometimes that clinical aspects of the trial (noncompliance, lower-than expected tolerated doses) suggest power less than desired against clinically meaningful alternatives. For this and other, ethical, reasons, there has been a perceived need for *adaptive* (group-)

sequential trial designs accommodating flexibility in accrual rates and spacing of look times. However, the designs must explicitly take account of such flexibility: Proschan et al. (1992) nicely illustrate the effects on significance level of modifying look-time definitions and other trial assumptions in mid-trial.

Up to the present, alternative suggestions for interim trial modifications are being proposed with optimal features of various sorts. A recent example is Case and Morgan (2001), which restricted itself to two-look two-armed trials with exponentially distributed survival but allowed accrual rates to be modfied at the interim look.

This paper has three objectives: first, to describe a Bayesian decision problem as in Leifer and Slud (2002) which incorporates multi-look trials with general loss-components penalizing trial length and incorrect decisions as a function of the alternative-parameter θ; second, to describe following Leifer and Slud (2002), especially in the two-look case, how optimal decision procedures require later look-times and stopping-boundaries to depend on earlier observed statistic values; and third, to describe an extended multi-look setting which includes the method Koutsoukos et al. (2000) used to design trials with an option to stop accrual with or without early stopping of the trial.

31.2 Decision Theoretic Formulation

In light of the theoretical results (Tsiatis 1982, Slud 1984) mentioned above, the data arising by calculating a two-sample (weighted-)logrank statistic at interim looks of a multi-look staggered-accrual trial with survival endpoints under local proportional-hazard alternatives (and also more general classes of alternatives), possibly requiring estimation of the variance increments in real time, is asymptotically equivalent in large data samples to the values of a Wiener process with drift, $X(t) = W_0(t) + \vartheta\, t$. Here ϑ is an unknown real parameter quantifying positive or negative relative prognosis for treatment- versus control-group patients in the trial. The objective of the trial is inference on ϑ to distinguish the null hypothesis $\vartheta \le 0$ against alternatives with $\vartheta > 0$: process data $X(\tau_j)$ may be observed at an increasing sequence of times τ_j , $1 \le j \le K$, with τ_j allowed to be determined from $(\tau_i, X(\tau_i),\ i < j)$ (and, possibly, auxiliary randomizations independent of the data). The upper-bound K on the number of look times is nonrandom and fixed, and the trial ends at the first time τ_ν for which either $\nu = K$ or $\tau_{\nu+1} = \tau_\nu$, at which time a binary decision $\chi \in \{0, 1\}$ is made as a function of all observable data $(\tau_i,\ X(\tau_i),\ i \le \nu)$. When actions $(\tau_i,\ 1 \le i \le \nu)$ and χ have been taken, losses are measured in terms of $\tau_\nu = t$ and $\chi = z \in \{0, 1\}$, when ϑ is the correct alternative (drift)

parameter assumed distributed according to a prior distribution π on $\mathbf{R}$, by

$$L(t,z,\vartheta) \;=\; \begin{cases} c_1(t,z,\vartheta) + z\, c_2(t,\vartheta) + (1-z)\, c_3(t,z,\vartheta), & \text{if} \quad \vartheta \le 0\,, \\ c_(t,z,\vartheta) + (1-z)\, c_2(t,\vartheta) + z\, c_3(t,z,\vartheta), & \text{if} \quad \vartheta > 0. \end{cases} \tag{31.2.2}$$

Note that z is the indicator of rejection of the null hypothesis $\mathbf{H}_0 : \vartheta \le 0$. The functions c_1, c_2, and c_3 represent, respectively, the costs of trial duration; incorrect terminal decision; and correct, but late, terminal decision. These costs are general enough to apply to realistic clinical trial scenarios, both from the point of view of public health and of drug developers. The cost functions are assumed to be π-integrable for each (t,z), piecewise smooth jointly in (t,ϑ), nondecreasing in t, and to satisfy for all (t,z,ϑ) :

$$c_1(0,\vartheta) = c_3(0,\vartheta) = 0 \quad , \qquad c_3(t,z,\vartheta) \;<\; c_2(t,\vartheta) \tag{31.2.3}$$

In addition, π is assumed to place positive mass in small neighborhoods of $\vartheta = 0$ and $\vartheta = \vartheta_1 > 0$, and $c_1(\cdot,z,\vartheta)$ is assumed to grow to ∞ for $z = 0, 1$ and π-almost all ϑ.

In this setting, the decision problem is to choose decision rules

$$\delta = (\{\tau_j\}_{j=1}^K,\; \nu,\; \chi) \tag{31.2.4}$$

subject for fixed $\alpha, \beta > 0$ to the type I and II error probability constraints

$$E_{\vartheta=0}(\chi) \le \alpha \quad , \qquad E_{\vartheta=\vartheta_1}(1-\chi) \le \beta \tag{31.2.5}$$

This decision theoretic problem closely mirrors that of Leifer and Slud (2002). It can be analyzed, standardly, in terms of a Lagrangian formulation (Berger 1985) in which the constraints (31.2.5) are omitted and the loss-function is replaced (after a reduction showing there is no loss of generality in assuming $\pi_0 \equiv \pi(\{0\}) > 0$ and $\pi_1 \equiv \pi(\{\vartheta_1\}) > 0$) by

$$L_{\lambda_0,\lambda_1}(t,z,\vartheta) \;\equiv\; L(t,z,\vartheta) \;+\; \frac{\lambda_0}{\pi_0}\, I_{[\vartheta=0]} \;+\; \frac{\lambda_1}{\pi_1}\, I_{[\vartheta=\vartheta_1]} \tag{31.2.6}$$

31.3 Optimal Decision Rules

Leifer and Slud (2002, including later revisions) show that optimal Bayesian decision rules for (31.2.6) have the following properties:

1. There is a finite, nonrandom constant $t_* > 0$, which may be made uniform with respect to compact sets of pairs $(\lambda_0, \lambda_1) \in \mathbf{R}^2_+$, such that $\tau_\nu \le t_*$.

2. For each triple (α, β, r) lying on the (closed) lower boundary of the 3-dimensional convex set of triples

$$\Big(E_{\vartheta=0}(\chi),\ E_{\vartheta=\vartheta_1}(1-\chi),\ \int E_\vartheta(L(\tau_\nu, \chi, \vartheta))\,\pi(d\vartheta)\Big) \tag{31.3.7}$$

of randomized decision rules, there exists a possibly randomized decision rule for which (α, β, r) is exactly equal to the triple (31.3.7).

3. Every minimum-risk, possibly randomized, decision rule for the decision problem with the loss-function (31.2.6) has a terminal decision χ which is a.s. equal to a nonrandom function of the form $\chi = I_{[X(\tau_\nu)\geq w(\tau_\nu)]}$, with $w(\cdot)$ unique, but implicitly defined.

4. Generically for the loss-function (31.2.6), possibly after a small random perturbation of the cost-function c_1 preserving the Assumptions, for each (λ_0, λ_1), the optimal decision rule minimizing the Bayesian risk for loss function (31.2.6) is unique and nonrandomized and can be computed by backward induction.

31.4 Modified Trials with Accrual-Stopping

We conclude by indicating (a one-sided modification of) a trial design of Koutsoukos et al. (2000) extending that of Section 31.2, which allows the flexibility of modifying accrual without stopping followup, effectively reducing, but not to 0, the rate at which information about the alternative parameter unfolds. The notation concerning the repeatedly calculated statistic S_k is as in the Introduction. In this design, the look-times $\tau_j = j$ are evenly spaced, since at most one unit of further followup is allowed when accrual is stopped, and at the end of such a followup period the trial is stopped. In the one-sided version of this design, trial and accrual stopping are respectively determined by a constant C_U and sequences $C_{A,j}$ and $C_{R,j}$ such that $C_{A,j} < C_U$:

> The trial is stopped outright at j, with Rejection, if $S_j/\sqrt{n} \geq C_U$.
>
> The accrual (i.e. entry) of new patients is disallowed for all times later than j if $C_{A,j} \leq S_j/\sqrt{n} < C_U$, in which case the trial is stopped at time $j+1$, with final Rejection if $S_{j+1}/\sqrt{n} \geq C_{R,j+1}$ and Acceptance otherwise.

Boundaries of this type can be computed to have fixed size, and the free parameters C_U, $C_{A,j}$, $C_{R,j+1}$ can be optimized with respect to a loss function containing costs for wrong decisions and trial durations (analogous to cost function c_1 in Section 31.2) under a range of alternatives weighted by a prior π.

We will indicate how this design falls within a slightly generalized version of the setting in Section 31.2 when the variance-process for the observed statistic process is affected by design elements, like the stopping of accrual at an interim look without stopping the trial.

References

1. Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

2. Case, L. and Morgan, T. (2001) Duration of accrual and follow-up for two-stage clinical trials, *Lifetime Data Analysis*, **7**, 21-37.

3. Fleming, T., Harrington, D. and O'Brien, P. (1984), Designs for group sequential tests, *Controlled Clinical Trials*, **5**, 348-361.

4. Jennison, C. (1987), Efficient group sequential tests with unpredictable group sizes, *Biometrika*, **77**, 577-513.

5. Koutsoukos, A., Rubinstein, L. and Slud, E. (2000), Early accrual-stopping sequential designs for clinical trials. *US National Cancer Institute*, preprint.

6. Lan, G. and DeMets, D. (1983) Discrete sequential boundaries for clinical trials, *Biometrika*, **70**, 659-663.

7. Leifer, E. and Slud, E. (2002), Optimal time-adaptive repeated significance tests. Preprint, under revision.

8. Proschan, M., Follmann, D. and Waclawiw, M. (1992), Effects of assumption violations on Type I error rate in group sequential monitoring, *Biometrics*, **51**, 1315-1324.

9. Slud, E. (1984) Sequential linear rank tests for two-sample censored survival data, *Annals of Statistics*, **12**, 551-571.

10. Slud, E. and Wei, L. J. (1982), Two-sample repeated significance tests based on the modified Wilcoxon statistic, *Journal of the American Statistical Association*, **77**, 862-868.

11. Tsiatis, A. (1982), Repeated significance testing for a general class of statistics used in censored survival analysis, *Journal of the American Statistical Association*, **77**, 855-861.

32

Breast Cancer Among Asian-American Women in Los Angeles County

Anna H. Wu

Keck School of Medicine
University of Southern California

32.1 Extended Abstract

Breast cancer is a disease that exhibits substantial international variation in incidence. Rates tend to be highest among whites in the United States and western Europe, whereas rates in Asia are among the lowest. Historically there existed about a 6-fold difference in breast cancer incidence between these extremes (Ziegler et al., 1993). This large variation in breast cancer risk is not due to underlying genetic differences as the rates of breast cancer in Asian-Americans shift substantially towards those of whites in the United States (Deapen et al., 2002).

To better understand reasons for the increase of breast cancer in Asian-Americans, we have conducted a large population-based case-control study which included 1,277 women with breast cancer and 1,160 control women without breast cancer. Chinese, Japanese and Filipino women, between the ages of 25 and 74 years at the time of diagnosis of an incident breast cancer on or after January 1995 were identified through the Los Angeles County Cancer Surveillance Program, the population-based cancer registry of the study area. Control women were frequency-matched to cases on specific Asian ethnicity and age and were selected from the neighborhoods where breast cancer cases resided at the time of diagnosis. In-person interviews were conducted by using a standardized, structured questionnaire that covered demographic characteristics and migration history, menstrual and reproductive history, lifetime use of exogenous hormones, body size at each decade of life, physical activity patterns, and diet history. The food frequency questionnaire was modeled after the validated diet instrument used in the Multiethnic Cohort Study being conducted

in Hawaii and Los Angeles County (Stram et al., 2000).

We calculated odds ratios (relative risk estimates), their corresponding 95% confidence intervals and P values by conditional logistic regression methods, with matched sets defined jointly by age, and specific Asian ethnicity. All basic regression models in this study also included as covariates birthplace and years of residence in the United Status, education, and relevant menstrual and reproductive factors. Results based on the first group of cases and controls interviewed showed that breast cancer risk was significantly inversely associated with soy intake and that age at first exposure to soy intake was an important co-determinant of protection (Wu et al., 2002). We also observed that tea intake, particularly green tea intake, had significant protective effect against breast cancer risk in Asian American women (Wu et al., 2003). Our updated findings from this study, including results in relation to menstrual and reproductive factors, dietary factors, body size, and use of exogenous hormones will be discussed. Some of the differences and similarities between the Asian diet and the Mediterranean diet will be highlighted.

We have recently expanded the scope of this study and are following the cases interviewed in this case-control study to determine the extent to which prediagnostic dietary and non-dietary lifestyle factors are associated with breast cancer prognosis in Asian-American women. A secondary goal is to explore if prognosis is associated with post-diagnostic lifestyle factors.

Data collected in the completed interview for the case-control study will be the source of information on exposures before diagnosis. We are recontacting cases by telephone to conduct a follow-up interview five or more years after initial cancer diagnosis. The follow-up interviews include questions on disease status (recurrences, new primaries), treatment history (surgery, use of tamoxifen, aromatase inhibitors and other agents, chemotherapy, radiation), use of alternative therapies, and selected lifestyle factors after diagnosis, including pregnancy, changes in menopausal status, use of exogenous hormones, body weight, physical activity, intake of soy and tea and other food groups. The Los Angeles County Cancer Surveillance Program will serve as the source of information on survival and tumor characteristics (tumor stage, nodal status, tumor size, histology, grade and estrogen/progesterone receptor status). Two endpoints will be included in the final analysis: overall survival and recurrence/second primary cancer. Some of our experience from this ongoing study will be discussed.

References

1. Deapen D, Liu L, Perkins C, Bernstein L, Ross RK (2002). Rapidly rising breast cancer incidence rates among Asian-American women. *Int J Cancer* **99**, 747-750.

2. Stram DO, Hankin JH, Wilkens LR, Pike MC, Monroe KR, Park S, Henderson B, Nomura AMY, Earle ME, Nagamine FS, Kolonel LN (2000). Calibration of the dietary questionnaire for a multiethnic cohort in Hawaii and Los Angeles. *Am J Epidemiol* **151**, 358-370.

3. Wu AH, Wan P, Hankin J, Tseng CC, Yu MC, Pike MC (2002). Adolescent and adult soy intake and risk of breast cancer in Asian-Americans. *Carcinogenesis* **23**, 1491-1496.

4. Wu AH, Yu MC, Tseng CC, Hankin J, Pike MC (2003). Green tea and risk of breast cancer in Asian-Americans. *Int J Cancer* **106**, 574-579.

5. Ziegler RG, Hoover RN, Pike MC, Hildsheim A, Nomura AMY, West DW, Wu-Williams AH, Kolonel LN, Horn-Ross PL, Rosenthal FJ, Hyer MB (1993). Migration patterns and breast cancer risk in Asian-American women. *J Natl Cancer Inst* **85**, 1819-1827.

33

Entropy and Divergence Measures for Mixed Variables

Konstantinos Zografos

University of Ioannina, Department of Mathematics, Ioannina, Greece & University of Cyprus, Department of Mathematics and Statistics, Nicosia, Cyprus

Abstract: It is well known the role of entropies and divergences in statistics and related fields as indices of the diversity or variability and as pseudo distances between statistical populations. The definition of these measures is extended in the case of mixed continuous and categorical variables, a case which is common in practice in the fields of medicine, behavioural sciences etc. The role of these indices in testing statistical hypothesis and as descriptive measures in the location model will be clarified.

Keywords and phrases: Location model, mixed variables, entropy, divergence

33.1 Introduction

Many times in practice the statistician is faced with mixed, continuous and categorical variables. In medicine, for instance, variables sex, profession, smoking and drinking are categorical while variables age, weight, height and time per week for gymnastic are continuous. In this and similar situations, the vector random variables include both, continuous and categorical components. There are several options to treat mixed data. If, for example, the qualitative variables can be subjected to some scoring system, then all variables can be treated as quantitative. In a similar manner, all the variables can be treated as qualitative if the quantitative variables might be categorized by grouping. Another approach is to analyze separately the continuous and the categorical part of the data and then to combine the results. But all of the above procedures involve, according to Krzanowski (1983), some element of subjectivity. If, for example, we treat the continuous variables as categorical by grouping them, then

this procedure results a loss of information due to the grouping of the observations. If we treat separately the continuous and the categorical variables and combine the results of the individual analyses then we will ignore possible associations and dependencies between the continuous and the categorical variables which may cause a false final decision. These reasons motivated several authors to adopt the location model, introduced by Olkin and Tate (1961) (cf. also Schafer (1997)), to study this type of mixed data. The location model helps to handle the joint distribution of mixed continuous and categorical variables and has been used to formulate statistical tests, as well as, discrimination and classification rules. Representative work in testing statistical hypothesis with mixed data are the papers by Afifi and Elashoff (1969), Bar-Hen and Daudin (1995), Morales *et al.* (1998), de Leon and Carrière (2000) and Nakanishi (2003). Allocation rules on this model were investigated, among other, by Krzanowski (1975), Vlachonikolis (1985), Balakrishnan *et al.* (1986), Nakanishi (1996).

On the other hand, information theoretic procedures are well known in statistics and related fields, and entropy and divergence measures provide with useful tools in order to formulate and define statistical tests and allocation rules. The use of entropies and divergences in the case of mixed variables is the subject of the papers by Krzanowski (1983), and recently by Bar-Hen and Daudin (1995), Morales *et al.* (1998) and Nakanishi (1996, 2003). To handle the joint distribution of mixed continuous and categorical variables, Krzanowski (1983) has considered the location model as it is introduced by Olkin and Tate (1961), while Bar-Hen and Daudin (1995) considered a generalization of the location model.

In this talk, some preliminary concepts will be presented in respect to the location model. This model will be applied in order to present measures of entropy and divergence in the mixed variables case. In this context, some probabilistic and statistical results will be outlined and discussed.

33.2 The Model

The location model has been introduced by Olkin and Tate (1961) and has since been used in several disciplines in statistics and related fields. In order to state this model consider q continuous random variables $X_1, ..., X_q$ and d categorical random variables $Y_1, ..., Y_d$, where each Y_i is observed at k_i, $i = 1, ..., d$, possible states y_{ij}, $i = 1, ..., d$ and $j = 1, ..., k_i$. The d qualitative random variables define a multinomial vector $Z = (z_1, ..., z_c)^T$ with c possible states, where each of the $c = k_1 \times k_2 \times ... \times k_d$ states is associated with a combination of the values y_{ij}, $i = 1, ..., d$ and $j = 1, ..., k_i$ of the qualitative variables. Denote by p_m the

probability of observing the state z_m, $m = 1, ..., c$,

$$p_m = \Pr(Z = z_m),\ m = 1, ..., c,\ \text{ with } \sum_{m=1}^{c} p_m = 1.$$

Conditionally on $Z = z_m$, $m = 1, ..., c$, the q continuous random variables $X = (X_1, ..., X_q)^T$ are described by a parametric density denoted by $f_m(x)$, that is,

$$f_m(x) = f(x|Z = z_m),$$

for $m = 1, ..., c$. In this context, the joint density with parameter θ of the random vectors $X = (X_1, ..., X_q)^T$ and $Z = (z_1, ..., z_c)^T$ is

$$\begin{array}{ll} f_\theta(x, z) & = \sum_{m=1}^{c} f(x|Z = z_m) \Pr(Z = z_m) I_{z_m}(Z) \\ & = \sum_{m=1}^{c} f_m(x) p_m I_{z_m}(Z), \end{array} \tag{33.2.1}$$

with

$$I_{z_m}(Z) = \left\{ \begin{array}{ll} 1, & \text{if } Z = z_m \\ 0, & \text{otherwise} \end{array} \right. ,\ \text{for } m = 1, ..., c.$$

The joint density $f_\theta(x, z)$, given by (33.2.1), defines the well known location model. The conditional density can be any parametric family of probability distributions. The classic location model, defined by Olkin and Tate (1961), considers that conditionally on $Z = z_m$, $m = 1, ..., c$, the q continuous random variables $X = (X_1, ..., X_q)^T$ jointly follow the multivariate normal distribution with location and scale parameters respectively μ_m and Σ_m, with Σ_m a positive definite matrix of order q, for $m = 1, ..., c$. If we will denote by $f_m(x)$ this conditional density, then

$$f_m(x) = f(x|Z = z_m) = (2\pi)^{-\frac{q}{2}} |\Sigma_m|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x - \mu_m)^T \Sigma_m^{-1} (x - \mu_m) \right\}. \tag{33.2.2}$$

Bar-Hen and Daudin (1995), generalized the classic location model (33.2.1) and (33.2.2) by considering $f_m(x) = f(x|Z = z_m)$ to be any parametric family of probability distributions and not necessarily the multivariate normal model (33.2.2).

Let μ_1 be the countable measure on $Z = \{z_1, \ldots, z_c\}$ and μ_2 be the Lebesque measure on R^q. Denote by $\mu = \mu_1 \otimes \mu_2$ the product measure on $Z \times R^q$. Then, for a concave function φ, the $\varphi-$entropy of the joint density $f_\theta(x, z)$, given by (33.2.1), is defined by

$$H_\varphi(f_\theta) = \int \varphi(f_\theta(x, z)) d\mu, \tag{33.2.3}$$

and it can be considered as a descriptive measure of the variability or diversity of the mixed variables and hence of their joint distribution. If $\varphi(x) = -x \ln x$, $x > 0$, then (33.2.3) leads to the well known Shannon entropy.

Suppose now that the continuous and the categorical variables $X_1, ..., X_q$ and $Y_1, ..., Y_d$, are observed on the members of two populations π_1 and π_2. Each of the populations is described by the generalized location model (33.2.1) with joint density

$$f_{\theta_i}(x, z) = \sum_{m=1}^{c} f_i(x|Z = z_m) p_{im} I_{z_m}(Z), i = 1, 2,$$

respectively, where p_{im} denotes the probability of observing the state z_m in the population π_i, $i = 1, 2$ and $m = 1, ..., c$.

The $\phi-$divergence of f_{θ_1} and f_{θ_2} is defined by

$$D_\phi(f_{\theta_1}, f_{\theta_2}) = \int f_{\theta_2}(x, z) \phi \left(\frac{f_{\theta_1}(x, z)}{f_{\theta_2}(x, z)} \right) d\mu, \tag{33.2.4}$$

where ϕ is a real convex function defined on $(0, \infty)$, which, moreover, satisfies appropriate conditions which ensure the existence of the above integral. Special choices of the convex function ϕ lead to the Kullback–Leibler directed divergence, the Cressie and Read's power divergence and the distances considered by Krzanowski (1983), as well. D_ϕ is a measure of the distance between populations π_1 and π_2 in the sense that $D_\phi(f_{\theta_1}, f_{\theta_2})$ attains its minimum value $\phi(1)$ if and only if $f_{\theta_1}(x, z) = f_{\theta_2}(x, z)$.

In practice, training samples are available from the population described by $f_\theta(x, z)$ or the populations π_1, π_2 and based on the samples, we are interested in the study of the sampling behavior of $H_\varphi(f_\theta)$, or to test the hypothesis of homogeneity of the two populations or to construct minimum distance rules for the allocation of a new observation as coming from one of the populations considered. In these cases an estimator of $H_\varphi(f_\theta)$ or $D_\phi(f_{\theta_1}, f_{\theta_2})$ can serve as a test statistic for testing homogeneity or as the main tool in order to define a minimum distance allocation rule. An estimator of $H_\varphi(f_\theta)$ or $D_\phi(f_{\theta_1}, f_{\theta_2})$ can be obtained, on the basis of a random sample of size n from $f_\theta(x, z)$, or on the basis of two independent random samples of sizes n_i, from the populations $f_{\theta_i}(x, z)$, $i = 1, 2$. Let $\widehat{\theta}$ denotes the m.l.e. of θ and $\widehat{\theta}_i$ the m.l.e. of θ_i, $i = 1, 2$. Then, the sample estimators of H_φ and D_ϕ are obtained from (33.2.3) and (33.2.4), if we replace the unknown parameters by their m.l.e., in the formulas for $H_\varphi(f_\theta)$ and $D_\phi(f_{\theta_1}, f_{\theta_2})$. The said estimators are the $\varphi-$entropy of $f_{\widehat{\theta}}$ and the $\varphi-$divergence of $f_{\widehat{\theta}_1}$ and $f_{\widehat{\theta}_2}$, defined, respectively, by

$$H_\varphi(f_{\widehat{\theta}}) = \int \varphi(f_{\widehat{\theta}}(x, z)) d\mu, \quad \text{and} \quad D_\phi(f_{\widehat{\theta}_1}, f_{\widehat{\theta}_2}) = \int f_{\widehat{\theta}_2}(x, z) \phi \left(\frac{f_{\widehat{\theta}_1}(x, z)}{f_{\widehat{\theta}_2}(x, z)} \right) d\mu.$$

In this talk, we will mainly concentrate, without any loss of generality, on the well known Shannon entropy and Kullback-Leibler divergence for mixed, continuous and categorical variables which are obtained from (33.2.3) and (33.2.4),

for $\varphi(x) = -x \ln x$ and $\phi(x) = x \ln x$, respectively. Asymptotic distributions of these entropy and divergence measures will be stated. Moreover, the role of the above indices as descriptive measures in the location model will be discussed and studied.

References

1. Afifi, A. A. and Elashoff, R. M. (1969). Multivariate two sample tests wih dichotomous and continuous variables. I. The location model. *Ann. Math. Statist.*, **40,** 290–298.

2. Balakrishnan, N., Kocherlakota, S. and Kocherlakota, K. (1986). On the errors of misclassification based on dichotomous and normal variables. *Ann. Inst. Statist. Math.*, **38**, 529-538.

3. Bar-Hen, A. and Daudin, J.-J. (1995). Generalization of the Mahalanobis distance in the mixed case. *J. Multivariate Anal.*, **53**, 332-342.

4. Daudin, J. J, and Bar-Hen, A. (1999). Selection in discriminant analysis with continuous and discrete variables. *Comput. Statist. Data Anal.*, **32**, 161-175.

5. de Leon, A. R. and Carriere, K. C. (2000). On the one sample location hypothesis for mixed bivariate data. *Commun. Statist. Theory Methods.*, **29**, 2573-2561.

6. Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Assoc.*, **70**, 782-790.

7. Krzanowski, W. J. (1983) . Distance between populations using mixed continuous and categorical variables. *Biometrika*, **70**, 235-243.

8. Morales, D., Pardo, L. and Zografos, K. (1998). Informational distances and related statistics in mixed continuous and categorical variables. *J. Statist. Plann. Inference,* **75**, 47–63.

9. Nakanishi, H. (1996). Distance between populations in a mixture of categorical and continuous variables. *J. Japan Statist. Soc.*, **26**, 221–230.

10. Nakanishi, H. (2003). Tests of hypotheses for the distance between populations on the mixture of categorical and continuous variables. *J. Japanese Soc. Comput. Statist.*, **16**, 53–62.

11. Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variable. *Ann. Math. Stat.*, **32**, 448-465.

12. Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall, New York.

13. Vlachonikolis, I. G. (1985). On the asymptotic distribution of the location linear discriminant function. *J. Roy. Statist. Soc.* Ser. B, **47**, 498–509.

Part III

Contributed Papers

34

The Power-Generalized Weibull Probability Distribution And Its Use In Survival Analysis

Roza Alloyarova*, Mikhail Nikulin, Natalie Pya*, and Vassilly Voinov***

**Kazakhstan Institute of Management, Economics and Strategic Research, Almaty, Kazakhstan*
***EA 2961. Statistique Mathematique Université Bordeaux 2, Bordeaux, France; V.A. Steklov Mathematical Institute, St Petersburg, Russia*

Abstract: Modified chi-squared tests for the Power-Generalized Weibull Probability Distribution are introduced. Power of those tests is investigated.

Keywords and phrases: maximum likelihood and moment type estimators; three-parameter Weibull, generalized Weibull, and the power generalized Weibull probability distributions; hypotheses testing; reliability; survival analysis; modified chi-squared test

34.1 Introduction

The Weibull probability distribution function $W(t;\theta,\nu)$ and its numerous generalizations are often used as a lifetime model (see, e.g., Bagdonavičius and Nikulin (2002), Barlow and Proschan (1991), Harter (1991)). A random nonnegative absolutely continuous failure time T can be described by the different ways. For example, the Weibull distribution of T can be specified by its probability distribution function

$$W(t;\theta,\nu) = 1 - \exp\{1 - (t/\theta)^{\nu}\}, \quad \theta,\nu,t > 0, \tag{34.1.1}$$

by the survival function $S(t;\theta,\nu) = P(T \geq t) = 1 - W(t;\theta,\nu)$, by the hazard or failure rate function $\alpha(t;\theta,\nu) = \nu\theta^{-\nu}t^{\nu-1}$, by the probability density function $f(t;\theta,\nu) = \nu\theta^{-\nu}t^{\nu-1}\exp[-(t/\theta)^{\nu}]$ or by quantile function $t_p = \theta[-\ln(1-p)]^{1/\nu}$, $0 < p < 1$, (see Lawless (2003)).

Introduction of Weibull family $W(t;\theta,\nu)$ can be justified by the following two reasons:

Remark 34.1.1 *Let $T_1,...,T_n$ be a random sample such that*

$$P(T_i \leq t) = G(t;\theta,\nu), \quad i = 1,...,n, \quad \theta,\nu,t > 0,$$

where $G(t;\theta,\nu)$ *is a distribution function satisfying the conditions* $\lim_{t\downarrow 0} G(t;\theta,\nu) = \theta^{-\nu}t^{\nu}$, $G(t;\theta,\nu) = 0$ *if* $t \leq 0$ *for all fixed* θ *and* ν. *Then* $n^{1/n}T_{(1n)}$, *where* $T_{(1n)} = min(T_1,...,T_n)$ *is the first order statistic, converges in probability to* $W(t;\theta,\nu)$.

Remark 34.1.2 *Let* $T \sim W(t;\theta,\nu)$. *Then the statistic* $Z = \ln T$ *follows the well-known extreme-value probability distribution*

$$P(Z \leq z) = 1 - \exp\{1 - \exp[(z-\mu)/\sigma]\},$$

where $\mu = \ln\theta$ *and* $\sigma = 1/\nu > 0$. *This distribution is also often used in reliability studies.*

In accelerated life studies the Generalized power Weibull family with the distribution function

$$F(t;\theta,\nu,\gamma) = 1 - \exp\{1 - [1 + (t/\theta)^{\nu}]^{1/\gamma}\}, \quad t,\theta,\nu,\gamma > 0, \tag{34.1.2}$$

proves to be very useful (Bagdonavičius and Nikulin (2002)). The family (34.1.2) with all moments being finite possesses very nice properties. Dependent on values of parameters the hazard rate function $\alpha(t;\theta,\nu,\gamma) = \nu\gamma^{-1}\theta^{-\nu}[1+(t/\theta)^{\nu}]^{1/\gamma-1}$ can be constant, monotone increasing or decreasing, $\cap$-shaped and $\cup$-shaped. Note also that $F(t;\theta,\nu,1) = W(t;\theta,\nu)$ and $F(t;\theta,1,1) = \mathcal{E}(t;\theta)$, which is the exponential probability distribution.

In Section 34.2 of this note we consider the problem of estimating parameters for the family (34.1.2). Section 34.3 is devoted to constructing a modified chi-squared test based on moment type estimators. Results of power estimating are analyzed in Section 34.4.

34.2 Estimating parameters

Let $T_1,...,T_n$ be a random sample from the distribution (34.1.2). Assuming parameters θ,ν,γ to be unknown, the loglikelihood function for (34.1.2) can hardly be maximized analytically. Because of this we investigated maximum likelihood estimates (MLEs) $\hat{\theta},\hat{\nu},\hat{\gamma}$ of parameters θ,ν,γ by Monte Carlo simulation. Results of the simulation showed that for three-parameter Power-Generalized Weibull probability distribution (34.1.2) MLEs do not converge to their true values, and, hence, are inconsistent.

Moments of the family (34.1.2) can be represented as

$$ET^j = \theta^j \int_1^{\infty} (t^{\gamma}-1)^{j/\nu} e^{1-t}dt, \quad j = 1,2,... \tag{34.2.3}$$

Moment type estimates (MMEs) $\bar{\gamma}$ and $\bar{\nu}$ can be found by solving the equation

$$g(\gamma,\nu)=\left\{\int\limits_1^\infty (t^\gamma-1)^{2/\nu}e^{1-t}dt\Big[\int\limits_1^\infty (t^\gamma-1)^{1/\nu}e^{1-t}dt\Big]^{-2}-\overline{T^2}/(\bar{T})^2\right\}^2+$$

$$+\left\{\int\limits_1^\infty (t^\gamma-1)^{3/\nu}e^{1-t}dt\Big[\int\limits_1^\infty (t^\gamma-1)^{1/\nu}e^{1-t}dt\Big]^{-3}-\overline{T^3}/(\bar{T})^3\right\}^2, \tag{34.2.4}$$

where $\bar{T},\overline{T^2},\overline{T^3}$ are three first initial sample moments. MME $\bar{\theta}$ is then defined by the formula

$$\bar{\theta}=\bar{T}\Big[\int\limits_1^\infty (t^{\bar{\gamma}}-1)^{1/\bar{\nu}}e^{1-t}dt\Big]^{-1}.$$

Unfortunately, there is no exact solution of (34.2.4). Approximate solution, which minimizes $g(\gamma,\nu)$, gives inconsistent $\bar{\gamma}$ and not $\sqrt{n}$- consistent estimates $\bar{\nu}$ and $\bar{\theta}$.

If the parameter γ is considered to be fixed, MMEs $\bar{\nu}$ and $\bar{\theta}$ can be found by solving the following system of two equations:

$$\int\limits_1^\infty (t^\gamma-1)^{2/\nu}e^{1-t}dt\Big[\int\limits_1^\infty (t^\gamma-1)^{1/\nu}e^{1-t}dt\Big]^{-2}=\overline{X^2}/(\bar{X})^2 \tag{34.2.5}$$

and

$$\theta=(\bar{X})\Big[\int\limits_1^\infty (t^\gamma-1)^{1/\nu}e^{1-t}dt\Big]^{-1}. \tag{34.2.6}$$

Fig. 34.1 summarizes results of Monte Carlo simulation of MMEs $\bar{\theta}$ and $\bar{\nu}$. It can be seen that these MMEs are $\sqrt{n}$ - consistent. This suggests constructing a modified chi-squared test of Mirvaliev (2001) for testing the null hypothesis about the model (34.1.2).

34.3 Modified chi-squared test

Denote the unknown parameter $\vec{\theta}=(\theta_1,\theta_2)^T$, where $\theta_1=\theta$, $\theta_2=\nu$. Denote also $p_i(\vec{\theta})=\int\limits_{\triangle_i(\vec{\theta})} dF(t;\theta,\nu,\gamma)$, $i=1,...,r$, where $\triangle_i(\vec{\theta})$ are nonintersecting random equiprobable cells with borders $a_i(\vec{\theta})=\sigma\{[1-\ln(1-i/r)]^\gamma-1\}^{1/\nu}$, $i=0,1,...,r$, $a_0(\vec{\theta})=0$, $a_r(\vec{\theta})=\infty$. If we introduce the vector $V^n(\vec{\theta})$ of standardized grouped frequencies with components $v_i^n(\vec{\theta})=[np_i(\vec{\theta})]^{-1/2}(N_i^n-np_i(\vec{\theta}))$, then the standard Pearson's statistic $X^2(\vec{\theta})$ would be $X^2(\vec{\theta})=V^{(n)T}(\vec{\theta})V^{(n)}(\vec{\theta})$. If $\vec{\theta}=(\theta_1,\theta_2)^T$ is unknown and is replaced by MME $\bar{\theta}$, the modified Mirvaliev's test can be used instead of the Pearson's statistic, which in the limit will possess

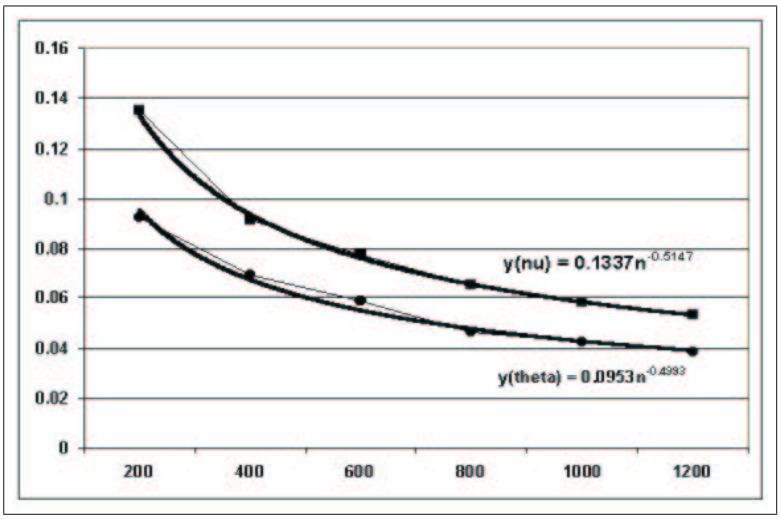


Figure 34.1: Simulated values of mean absolute errors for MMEs $\bar{\theta}, \bar{\nu}$ as a function of the sample size n ($\theta = \nu = 3, \gamma = 2$ fixed).

the chi-squared probability distribution with $r-1$ degrees of freedom and on the contrary of Pearson's test will not depend on unknown parameters. Define matrices $K(\vec{\theta}), V(\vec{\theta}), C(\vec{\theta})$, and $B(\vec{\theta})$ with elements $K_{ij} = \int\limits_0^\infty t^i \frac{\partial f(t;\theta,\nu,\gamma)}{\partial \theta_j} dt$, $V_{ij} = m_{ij} - m_i m_j$, where $m_i = E(T^i), m_{ij} = E(T^{i+j}), i,j = 1,2$, $C_{ij} = p_i^{-1/2}(\vec{\theta})\Big\{ \int\limits_{\triangle_i(\vec{\theta})} t^j f(t;\theta,\nu,\gamma)dt - p_i(\vec{\theta}) m_j \Big\}$, and $B_{ij} = \frac{1}{\sqrt{p_i(\vec{\theta})}} \int\limits_{a_{i-1}(\vec{\theta})}^{a_i(\vec{\theta})} \frac{\partial f(t;\theta,\nu,\gamma)}{\partial \theta_j}$, $i = 1,...,r, j = 1,2$, correspondingly. If $\vec{q} = (p_1^{1/2}(\vec{\theta}),...,p_r^{1/2}(\vec{\theta}))^T$ let matrices A and L be $A = I - \vec{q}\vec{q}^T + C(V - C^T C)^{-1} C^T$ and $L = V + (C - BK^{-1}V)^T A (C - BK^{-1}V)$ correspondingly. For the brevity we omitted dependence of all these matrices on $\vec{\theta}$. In terms of these matrices the Mirvaliev's statistic is written as (Mirvaliev (2001))

$$Y2^2(\bar{\theta}) = X^2(\vec{\theta}) + R^2(\bar{\theta}) - Q^2(\bar{\theta}), \tag{34.3.7}$$

where

$$R^2(\bar{\theta}) = V^{(n)T}(\bar{\theta}) C (V - C^T C)^{-1} C^T V^{(n)}(\bar{\theta}),$$

$$Q^2(\bar{\theta}) = V^{(n)T}(\bar{\theta}) A (C - BK^{-1}V) L^{-1} (C - BK^{-1}V)^T A V^{(n)}(\bar{\theta}).$$

The statistic $Y3^2(\bar{\theta}) = Y2^2(\bar{\theta}) - U^2(\bar{\theta})$, where $U^2(\bar{\theta}) = V^{(n)T}(\bar{\theta})[I - B(B^T B)^{-1} B^T] V^{(n)}(\bar{\theta})$ is the well known Dzhaparidze-Nikulin (DN) test (Dzhaparidze and Nikulin (1974)), which is distributed asymptotically as χ^2_2, can be used on its own right. Explicit expressions for elements of matrices A, B, C, K, L, and V will be published elsewhere.

34.4 Power estimating

To assess power of Mirvaliev $Y2^2(\bar{\theta})$ and $Y3^2(\bar{\theta})$ tests for the PGW null hypothesis against Exponentiated Weibull (EW) (see Mudholkar, Srivastava, and Freimer (1995)), Three-parameter Weibull (W3), and Generalized Weibull (GW) (see Mudholkar, Srivastava, and Kollia (1996)) alternatives we conducted Monte Carlo simulation for the different number r of equiprobable random cells. Anderson-Darling A^2 test was also simulated for comparison. We used samples of size $n = 200$, and type one error $\alpha = 0.05$. Table 34.1 summarizes results that were obtained. From the table one may conclude that shapes of PGW and

Table 34.1:

r	$Y2^2(\bar{\theta})$	$Y3^2(\bar{\theta})$	A^2
PGW-EW			
5	0.056	0.0510	0.052
15	0.049	0.061	0.052
40	0.063	0.0705	0.052
PGW-W3			
5	0.068	0.098	0.1205
15	0.060	0.077	0.1205
40	0.082	0.047	0.1205
PGW-GW			
5, 15, 40	1.000	1.000	1.000

EW distributions are very close to each other and no one test can definitely discriminate them. The same conclusion is true for testing PGW versus W3.

In the above investigation we tested a null hypothesis against an alternative assuming parameters being unknown. This means that parameters of the null hypothesis are adjusted to a sample generated by the alternative model thus making null and alternative hypotheses as close as possible. In other words words tests are sensitive only to the difference in shape of those hypotheses. From the results obtained it follows that shapes of the Generalized Power Weibull, Exponentiated Weibull, and Three-parameter Weibull distributions are very close to each other, though their hazard rate functions can be different. Thus, to select one of these models for a survival analysis we need to develop a test, which will compare their hazard rate functions directly (see also Voinov, Alloyarova, and Pya (2006)). At the same time to discriminate between different in their shape PGW and GW any test - $Y2^2(\bar{\theta})$, $Y3^2(\bar{\theta})$, and even insensitive DN $U^2(\bar{\theta})$ can be used, since power of all these tests is very

close to one.

References

1. Bagdonavičius, V. and Nikulin, M. (2002). *Accelerated Life Models: Modeling and Statistical Analysis*, Boca Raton: Chapman and Hall/CRC.

2. Barlow, R.E. and Proschan, F. (1991). Life Distribution Models and Incomplete Data, In *Handbook of Statistics 7* (Ed., P.R. Krishnaian and C.R. Rao), pp. 225–250.

3. Dzhaparidze, K.O. and Nikulin, M.S. (1974). On a modification of the standard statistic of Pearson, *Theory of Probability and its Applications*, **19**, 851–852.

4. Harter, H.L. (1991). Weibull, Log-Weibull and Gamma Order Statistics, In *Handbook of Statistics 7* (Ed., P.R. Krishnaian and C.R. Rao), pp. 433–466.

5. Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

6. Mirvaliev, M. (2001). An investigation of generalized chi-squared type statistics, *Doctoral thesis*, Academy of Science of the Republic of Uzbekistan, Tashkent.

7. Mudholkar, G.S., Srivastava, D.K., and Freimer, M. (1995). The exponentiated Weibull family: a reanalysis of the bus-motor-failure data, *Technometrics*, **37**, 436–445.

8. Mudholkar, G.S., Srivastava, D.K., and Kollia G.D. (1996). A generalization of the Weibull distribution with application to the analysis of survival data, *Journal of American Statistical Association* , **91**, 1575–1583.

9. Voinov, V.G., Alloyarova, R., and Pya, N. (2006). A Modified Chi-squared Test for the Three-parameter Weibull Distribution and its Applications in Reliability, (Submitted to the Int. Conf. on Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing, May 22-24, Angers, France).

35

On some Modification of Seemingly Unrelated Regression Equations Model

Alexander Andronov, Andrey Svirchenkov

Riga Technical University, 1 Kalku Str., LV-1658, Riga, Latvia.
e-mail lora@mailbox.riga.lv
Transport and Telecommunication Institute, 1 Lomonosov Str., LV-1019, Riga, Latvia. e-mail: secretary@lateko.lv

Abstract: The seemingly unrelated regression equations model considers some regression equations which are contemporaneously correlated. With that each observation gives the values of all equations, whereas in the proposed model this requirement is omitted.

Keywords and phrases: Regression equations, contemporaneous correlation.

35.1 Problem setting

We consider a group of G objects with numbers $i = 1, 2, ..., G$. The i-th object is examined n_i times, at the time moments $t_{i,1} < t_{i,2} < ... < t_{i,n_i}$. At the j-th time moment $t_{i,j}$ we fixe a vector of independent variables $x_{i,j} = (x_{i,j}^{(1)}, x_{i,j}^{(2)}, ..., x_{i,j}^{(m_i)})$ and a value of a dependent variable $Y_{i,j}$. It is supposed that the last is formed by the linear-regression equation

$$Y_{i,j} = \sum_{v=1}^{m_i} \beta_{i,v} x_{i,j}^{(v)} + Z_{i,j}, \tag{35.1.1}$$

where $\beta_{i,v}$ is the coefficient for the i-th object and v-th independent variable and $Z_{i,j}$ is normally distributed random term (a disturbance) with mean zero and variance σ_i^2.

Further if for two various objects i and i' the time moments $t_{i,j}$ and $t_{i',j'}$ coincide then the random terms $Z_{i,j}$ and $Z_{i',j'}$ (therefore $Y_{i,j}$ and $Y_{i',j'}$ too) are correlated random variables with the covariance $c_{i,i'}$ whereas for various time moments they are assumed independent ($Z_{i,j}$ and $Z_{i,j'}$ are independent for $j \neq j'$ as well).

As usually it is assumed that for $i = 1, 2, ..., G$ and $j = 1, 2, ..., n_i$ $x_{i,j} = (x_{i,j}^{(1)}, x_{i,j}^{(2)}, ..., x_{i,j}^{(m_i)})$ is known constant vector and $Y_{i,j}$ is the fixed value. On this base it should estimate the unknown parameters of the regression model $\{\beta_{i,v}\}$, $\{c_{i,i'}\}$, where $\sigma_i^2 = c_{i,i}$.

Our final aim is for the future time moment t to get a prognosis of the sum

$$W(t) = \sum_{i=1}^{G} Y_{i,j(t)} = \sum_{i=1}^{G} \Big(\sum_{v=1}^{m_i} \beta_{i,v} x_{i,j(t)}^{(v)} + Z_{i,j(t)} \Big) \tag{35.1.2}$$

The seemingly unrelated regression equations models were considered by Turkington (2002). In this case it is supposed that $n_i = n$, $t_{i,j} = t_{i',j}$ for all i, i'. If in addition $x_{i,j} = (x_{i,j}^{(1)}, x_{i,j}^{(2)}, ..., x_{i,j}^{(m_i)}) = x_{i',j} = (x_{i',j}^{(1)}, x_{i',j}^{(2)}, ..., x_{i',j}^{(m_i)})$ for all i, i', then we have the multivariate linear regression model, see Srivastava (2002).

35.2 The prognosis of the sum

Let X_i be a $(n_i \times m_i)$-matrix of the independent variables for the i-th object. The unknown coefficients $\{\beta_{i,v}\}$ are estimated using well known formula

$$\beta_i^* = (X_i^T X_i)^{-1} X_i^T Y_i, \tag{35.2.3}$$

where $\beta_i^* = (\beta_{i,1}^*, \beta_{i,2}^*, ..., \beta_{i,m_i}^*)^T$, $Y_i = (Y_{i,1}, Y_{i,2}, ..., Y_{i,n_i})^T$ are the vectors of the estimators and the dependent variables.

The last formula gives the unbiased estimator of β_i. It allows us to get the unbiased estimator for the sum of interest $W(t)$:

$$W^*(t) = \sum_{i=1}^{G} Y_{i,j(t)}^* = \sum_{i=1}^{G} x_{i,j(t)} \beta_i^* = \sum_{i=1}^{G} \sum_{v=1}^{m_i} \beta_{i,v}^* x_{i,j(t)}^{(v)}. \tag{35.2.4}$$

Our aim is the calculation of the variance of this estimator.

35.3 Covariance of the estimators

With respect to Eq.(35.2.3), the covariance matrix of two coefficient vectors β_i^*, β_l^* is calculated by formula

$$Cov(\beta_i^*, \beta_l^*) = (X_i^T X_i)^{-1} X_i^T Cov(Y_i, Y_l) X_l (X_l^T X_l)^{-1}. \tag{35.3.5}$$

Also we need to calculate the covariance $Cov(Y_i, Y_l)$. Let $D^{(i,l)}$ is $(n_i \times n_l)$-matrix for which $D_{j,f}^{(i,l)} = 1$ if $t_{i,j} = t_{l,f}$ and $D_{j,f}^{(i,l)} = 0$ otherwise.

Let us remember that the covariance of two dependent variables $Y_{i,j}$ and $Y_{l,f}$ for the same time moment $t_{i,j} = t_{l,f}$ is equal to $c_{i,l}$. Therefore

$$Cov(Y_i, Y_l) = c_{i,l} D^{(i,l)}.$$

Now we are able to rewrite Eq.(35.3.5) in the following form:

$$Cov(\beta_i^*, \beta_l^*) = c_{i,l}(X_i^T X_i)^{-1} X_i^T D^{(i,l)} X_l (X_l^T X_l)^{-1}. \tag{35.3.6}$$

Our next task is the estimation of the unknown covariance $\{c_{i,l}\}$. With that we try to use usual estimator of the least squares:

$$c_{i,l}^* = \frac{1}{v_{i,l}} (Y_i - X_i \beta_i^*)^T D^{(i,l)} (Y_l - X_l \beta_l^*), \tag{35.3.7}$$

where $v_{i,l}$ is a constant that is determined by a condition of the estimator unbiasedness.

To define the constant $v_{i,l}$ it is necessary to calculate expectation of the estimator $c_{i,l}^*$. We have:

$$E(c_{i,l}^*) = E\left(\frac{1}{v_{i,l}} (Y_i - X_i \beta_i^*)^T D^{(i,l)} (Y_l - X_l \beta_l^*)\right) =$$

$$= \frac{1}{v_{i,l}} E\left((Y_i - X_i (X_i^T X_i)^{-1} X_i^T Y_i)^T D^{(i,l)} (Y_l - X_l (X_l^T X_l)^{-1} X_l^T Y_l)\right) =$$

$$= \frac{1}{v_{i,l}} E\left(Y_i^T (I_i - X_i (X_i^T X_i)^{-1} X_i^T) D^{(i,l)} (I_l - X_l (X_l^T X_l)^{-1} X_l^T) Y_l\right),$$

where I_i and I_l are unique matrices of rank n_i and n_l respectively.

Since $Y_i = X_i \beta_i + Z_i$ then

$$(I_i - X_i (X_i^T X_i)^{-1} X_i^T Y_i = (I_i - X_i (X_i^T X_i)^{-1} X_i^T) Z_i,$$

$$E(c_{i,l}^*) = \frac{1}{v_{i,l}} E\left(Z_i^T (I_i - X_i (X_i^T X_i)^{-1} X_i^T) D^{(i,l)} (I_l - X_l (X_l^T X_l)^{-1} X_l^T) Z_l\right) \tag{35.3.8}$$

We introduce the following notation: $R_j(i)$ is the j-th row of the matrix $(I_i - X_i (X_i^T X_i)^{-1} X_i^T)$, $f(l,i,j)$ is the observation number for the l-th object, for which the time coincides with $t_{i,j}$, and is equal to zero if such number absents itself:

$$f(l,i,j) = \sum_{v=1}^{n_l} v \times D_{j,v}^{(i,l)}. \tag{35.3.9}$$

Then

$$E(c_{i,l}^*) = \frac{1}{v_{i,l}} \sum_j \sum_k E\left(Z_{i,j} R_j(i) D^{(i,l)} R_k^T(l) Z_{l,k}\right) =$$

$$= \frac{1}{v_{i,l}} \sum_j R_j(i) R^T_{f(l,i,j)}(l) E\left(Z_{i,j} Z_{l,f(l,i,j)}\right) = \frac{1}{v_{i,l}} c_{i,l} \sum_j R_j(i) R^T_{f(l,i,j)}(l).$$

The last formula shows that

$$v_{i,l} = \sum_j R_j(i) R^T_{f(l,i,j)}(l). \tag{35.3.10}$$

With this value of the constant $v_{i,l}$ Eq.(35.3.7) gives the unbiased estimator of the covariance $c_{i,l}$.

Note for $\sigma_i^2 = c_{i,i}$ we have the usual estimator:

$$\sigma_i^{2*} = c^*_{i,i} = \frac{1}{n_i - m_i} (Y_i - X_i \beta_i^*)^T (Y_i - X_i \beta_i^*). \tag{35.3.11}$$

35.4 Variance of the sum

The variance of the sum in Eq.(35.2.4) is calculated by the usual way:

$$Var(W^*(t)) = \sum_{i=1}^{G} x_{i,j(t)} Cov(\beta_i^*) x^T_{l,j(t)} +$$

$$+2 \sum_{i=1}^{G-1} \sum_{l=i+1}^{G} x_{i,j(t)} Cov(\beta_i^*, \beta_l^*) x^T_{l,j(t)}. \tag{35.4.12}$$

With that $Cov(\beta_i^*, \beta_l^*)$ is calculated by Eq.(35.3.6) and the covariance matrix of the vector β_i^* is calculated by the well known formula

$$Cov(\beta_i^*) = c_{i,i} (X_i^T X_i)^{-1}. \tag{35.4.13}$$

35.5 Example

Consider two objects ($G = 2$) with numbers $i = 1, 2$. The first object is examined $n_1 = 5$ times, at the time moments $t_{1,1} = 1$, $t_{1,2} = 2$, $t_{1,3} = 4$, $t_{1,4} = 6$, $t_{1,5} = 9$. The second object is examined $n_2 = 7$ times, at the time moments $t_{2,1} = 1$, $t_{2,2} = 3$, $t_{2,3} = 4$, $t_{2,4} = 5$, $t_{2,5} = 6$, $t_{2,6} = 7$, $t_{2,7} = 9$. It is supposed that the dependent variables $\{Y_{1,j}, Y_{2,j}\}$ are formed by the following linear-regression equations:

$$Y_{1,j} = \beta_{1,1} + \beta_{1,2} t_{1,j} + \beta_{1,3} t_{1,j}^2 + Z_{1,j}, j = 1, ..., 5,$$

$$Y_{2,j} = \beta_{2,1} + \beta_{2,2} t_{2,j} + \beta_{2,3} t_{2,j}^2 + \beta_{2,4} \frac{1}{t_{2,j}^2} + Z_{2,j}, j = 1, ..., 7.$$

Let us calculate unique normalized constant $v_{1,2}$ using Eq.(35.3.10). The matrices X_1, X_2 and $D^{(1,2)}$ have the following forms here:

$$X_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 9 & 81 \end{pmatrix}, \; X_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 1/9 \\ 1 & 4 & 16 & 1/16 \\ 1 & 5 & 25 & 1/25 \\ 1 & 6 & 36 & 1/36 \\ 1 & 7 & 49 & 1/49 \\ 1 & 9 & 81 & 1/81 \end{pmatrix}, \; D^{(1,2)T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Matrix $R_1 = (I_1 - X_1(X_1^T X_1)^{-1} X_1^T)$ for example is

$$R_1 = \begin{pmatrix} 0.283 & -0.409 & 0.014 & 0.177 & -0.065 \\ -0.409 & 0.660 & -0.212 & -0.096 & 0.057 \\ 0.014 & -0.212 & 0.539 & -0.440 & 0.099 \\ 0.177 & -0.096 & -0.440 & 0.484 & -0.126 \\ -0.065 & 0.057 & 0.099 & -0.126 & 0.034 \end{pmatrix}.$$

The function $f(2,1,j)$ values are given in table 1.

Table 1

The function $f(2,1,j)$ values

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$f(2,1,j)$	1	0	3	5	7

The calculations according to Eq.(35.3.10) show that $v_{1,2} = 0.491$. We wish to remark that the result is not the whole number as it takes place in the usual regression analysis.

Now we wish to compare variances estimators for two cases: 1) the random variables $Y_{1,j}$ and $Y_{2,j}$ are independent; 2) at the same time they are dependent.

For the first case we have with respect to Eq.(35.4.12) and Eq.(35.4.13):

$$Var(W^*(t)) = \sum_{i=1}^{2} x_{i,j(t)} Cov(\beta_i^*) x_{i,j(t)}^T = \sum_{i=1}^{2} \sigma_i^2 x_{i,j(t)} (X_i^T X_i)^{-1} x_{i,j(t)}^T \quad (35.5.14)$$

For the second case

$$Var(W^*(t)) = \sum_{i=1}^{2} \sigma_i^2 x_{i,j(t)} (X_i^T X_i)^{-1} x_{i,j(t)}^T + 2x_{1,j(t)} Cov(\beta_1^*, \beta_2^*) x_{2,j(t)}^T, \quad (35.5.15)$$

where covariance $Cov(\beta_i^*, \beta_l^*)$ is calculated by Eq.(35.3.6). Let $t = 10$, then

$$x_{1,j(10)} = (1\ 10\ 100), x_{2,j(10)} = (1\ 10\ 100\ 0.01).$$

Let $\sigma_1^2 = c_{1,1} = 2$, $\sigma_2^2 = c_{2,2} = 5$, $c_{1,2} = \rho\sigma_1\sigma_2$ where ρ is the correlation coefficient. Then for first case the Eq.(35.5.14) gives $Var(W^*(10)) = 16.703$. For the second case the values of this variance (Eq. (35.5.15)) are presented in the table 2 as a function of the correlation coefficient ρ.

Table 2

The variance $Var(W^*(10))$ values as a function of ρ (four joint observations)

ρ	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
Var	7.474	10.11	12.75	15.39	18.02	20.66	23.30	25.93	28.57

We see that the dependence changes the variance values very sufficiently. It can be due to big number of joint observations (four out of five for the variable Y_1). Obviously the less the number of the joint observations the less this dependence. The table 3 contains corresponding results for three joint observations when the variable Y_1 is fixed at the time moments $t_{1,1} = 1$, $t_{1,2} = 2$, $t_{1,3} = 4$, $t_{1,4} = 8$, $t_{1,5} = 9$.

Table 3

The variance $Var(W^*(10))$ values as a function of ρ (three joint observations)

ρ	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
Var	9.358	11.35	13.33	15.32	17.31	19.30	21.29	23.28	25.26

Finally we can conclude that it is necessary to attach great importance to considered phenomena of the dependence in the given statistical data.

References

1. Srivastava, M.S. (2002). *Methods of Multivariate Statistics.* John Wiley & Sons, Inc., New York.

2. Turkington, D.A. (2002). *Matrix Calculus and Zero-One Matrices: Statistical and Econometric Applications.* Cambridge University Press, Cambridge.

36

Critical Condition in Human. The Entropy Based Technology of Definition.

Antonov V., Fedulin A., Nosyrev S., Kovalenko A, Kashtanov A.

Saint-Petersburg State Technical University
Hospital 122

Abstract: Technology of definition allows to determine the state of system and get the forecast of a state transition in real-time mode on usage of periodic signals of a system of any nature. The technology details are discussed here based on definition of correlation dimension in measured signal entropy phase.

Keywords and phrases:deterministic chaos, informational entropy, correlation dimension

36.1 Introduction

The know-how based on usage of periodic signals of a system of any nature – human, mechanical, thermal and chemical. It allows to determine the state of system and get the forecast of a state transition in real-time mode.

The state of system is determined by the value of correlation dimension of signal informational entropy in time. It can be explained such as the ability of the system to react on external stress. This way, changing of reaction ability means state transition.

The details of the computational model of such system state analysis are discussed. The main features are:

- Real-time mode of computation
- The data insufficiency (express analysis)
- Input noise
- Portability for mobile & other platforms with low computational ability

36.2 Methods

We present fast and simple way to obtain result. The every step in time of dynamic process split in 3 independent stages.

36.2.1 Stage 1. Evaluation of general signal frequency & amplitude

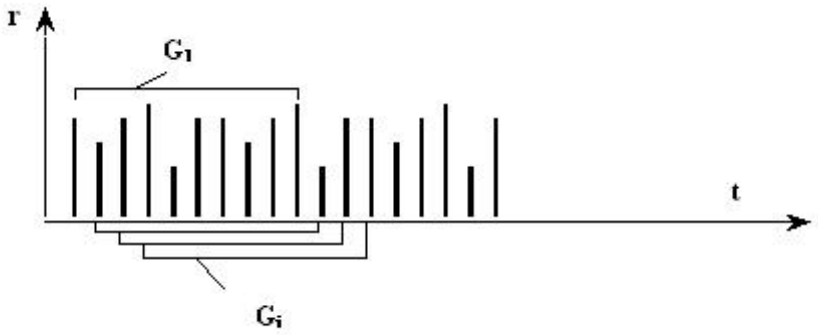
Let $X = \{x_i\}, i = 1..M_0$ – is source signal data, measured with frequency Δf. Let $R = \{r_i, t_i\}, i = 1..M_1$ – peak value and peak time.

To get peak values and peak time we can use Fast Fourier Transformation (36.2.1) or another well-known method from signal analysis theory.

$$\varphi(y) = \int_{-\infty}^{\infty} \varphi(x) e^{-ixy} dx \tag{36.2.1}$$

36.2.2 Stage 2. Evaluation of informational entropy

Let $G_i = \{r_{i-j}, t_{i-j}\}, j = 1..K$, series of peak values in time



For each G_i we calculate the value of informational entropy e in the following way [2]:

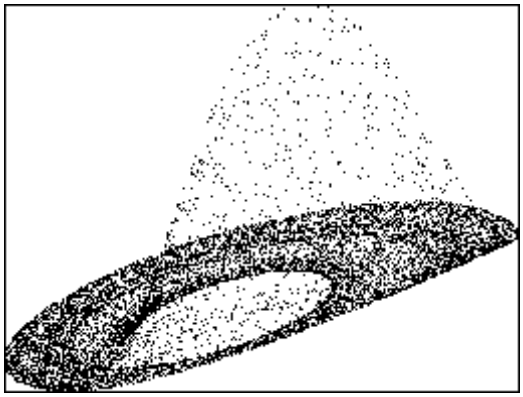
$$e = \sum_{i=1}^{n} P_i \log \frac{1}{P_i}, \tag{36.2.2}$$

where $\{P_i\}$– value distribution histogram of G_i.

To get the distribution histogram to calculate entropy value we can use many methods, for example, evaluation of autocorrelation function or straight evaluation of P_i in case where peak sampling frequency is much less than K. The main feature of the methods is that data value must be stable to K changing.

Let $E = \{e_i, t_i\}, i = 1..M_2$, system trajectory in entropy phase space, like it's shown on the figure below.

It's obvious to use the equation $M_1 = M_2 + K$ to get M_2. In fact Kdefines the value of time delay for getting first entropy value. The smaller the K value the faster we get the results, the greater the K value the more accuracy of calculation. The K value is determined by the entropy evaluation method.



36.2.3 Stage 3. Evaluation of correlation dimension of fractal curve in entropy phase space

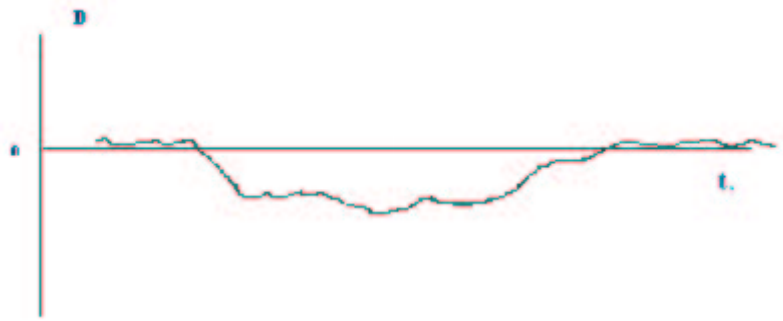
We suppose that the system trajectory E is the fractal [1] and evaluate correlation dimension of its attractor in time assuming its value is between 1 and 2. Correlation dimension of trajectory attractor is calculated by Hausdorf box method [4]. We subdivide all space in $M(\varepsilon)$ squares with length ε and calculate the probability of trajectory attractor to visit every square p_i. This way, correlation dimension is calculated by the formula

$$D_c^{=} \lim_{\varepsilon \to 0} \frac{\ln \sum_{i=1}^{M(\varepsilon)} p_i^2}{\ln \varepsilon} \tag{36.2.3}$$

To get value in range [1..2] we have to build $2D$ trajectory in $2D$ phase space $\vec{e_i}$=(x=value, y=value velocity).

Let $E_i = \{\vec{e_i} = (x = e_{i-j}, y = e_{i-j-L}), t = t_{i-j}\}$, j=1..N – part of system trajectory. It's proved by Tuckens [3] that it can be find L value that such repaired trajectory metrical features is the same with source built in phase space.

Let $D = \{d_i, t_i\}, i = 1, ..., M_3$, where d_i – value of correlation dimension of E_i. We can find M_3 from equation $M_3 = M_2 - L - N$. The L and N value defines the time delay value of first analysis results after input measurement



have been started, then each new d_i would be calculated after each new peak had been got.

In fact of data insufficiency we aren't able to use Hausdorf formula directly because this way all p_i are equal to 0. So we use the following method

$$C_c^(r) = \frac{1}{M(r)^2} \sum_{i,j=1}^{M(r)} H(r - |\vec{e}_i - \vec{e}_j|), \tag{36.2.4}$$

where H is Heavyside function. In other words $C_c(r)$ is the relative amount of point pairs closer in space to each other than r.

If $C_c(r) \cong \alpha r^D$ in some interval of r $R=(r_1,...,r_2)$ than we assume D_c=D– correlation dimension of trajectory attractor. To get D value we calculate series $\Delta C_c(r_i)$ with fixed logarithmic step by r:

$$\Delta C(r_i) = \frac{\log(C(r_{i+1}) - \log C(r_i))}{\log(r_{i+1}) - \log(r_i)} \tag{36.2.5}$$

It's evidently that $\Delta C_c(r_i = 0) = 0$ and $\Delta C_c(r_i{=}\infty){=}\ 1$, but if interval R exist then

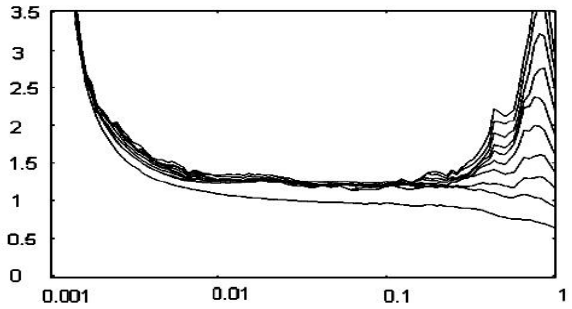
$$\Delta C(r_i) = \frac{\log C(r_{i+1}) - \log C(r_i)}{\log(r_{i+1}) - \log(r_i)} \approx \frac{\log(\frac{\alpha r_{i+1}^D}{\alpha r_i^D})}{\log(\frac{r_{i+1}}{r_i})} = D \tag{36.2.6}$$

so $\Delta C_c(r)$ is closer to constant function and function value defines D like it's shown on figures below.

It's shown by Tsonis [5] that the method convergence is good where $N{\geq}N_{min} = 10^{2+0.4D}$, so N value is determined from this equation. L value is determined from evaluation stability principle and can be from range $(1, ...N)$.

36.3 Summary

Given technology has been implemented as demonstration real-time application to determine human state using their ECG data (R-R peaks) power by Hospital



122. It's visible the value decreasing tendency for human with poor state of health.

References

1. Mandelbrot B. (1977). Fractals: Form, Chance, Dimension. Freeman, San-Francisco.

2. Farmer J.D. (1982). Information dimension and the probabilistic structure of chaos. *Z. Naturforsch.* **37**, 1304-1325.

3. Takens, F. (1985). On the numerical determination of the dimension of an attractor, In *Dynamical systems and bifurcations* (Eds. B.L.J. Braaksma, H.W. Broer and F. Takens). *Lect. Notes in Math.* **1125**, Springer, Heidelberg, 99-106.

4. Grassberger P., Procaccia I. (1983). Measuring the strangeness of strange attractors. *Physica D* **9**, 189-208 (1983).

5. Tsonis A. (1992). *Chaos: from Theory to Applications.* NY. Premium Press.

37

Analysis of Duration of Studies Data By Kernel Methods

Dimitrios Bagkavos and Aglaia Kalamatianou

Panteion University, Department of Sociology, 136 Syggrou Ave. 17675, Athens, Greece

Abstract: We use kernel based estimators to analyze duration of studies data. We consider separately female and male student groups as well the total population. The results are interpreted and illustrated graphically.

Keywords and phrases: Kernel estimate, censored data, duration of studies, survival function, hazard rate.

37.1 Introduction

Kalamatianou and McClean (2003) modeled the distribution of the duration of undergraduate studies in a Greek university using survival analysis techniques. In particular their nonparametric estimation part is based on the Kaplan-Meier estimator (Kaplan and Meier (1958)) which provides a step function as an estimate of the true survival function. Although the Kaplan-Meier curve is a well established method in survival analysis the information it provides is limited as it produces a step function. For this reason in this paper we work on the same problem and employ kernel estimates to obtain continuous curves of improved performance over the Kaplan-Meier estimate. Kulasekera et al. (2001) proved that a smoothed version of the Kaplan-Meier curve would be more efficient in terms of its asymptotic properties. This, together with the work of Marron and Padgett (1987) which extents kernel estimates to censored data situations yields continuous estimates of the survival and hazard functions appropriate for the data available at hand.

Next we give a brief description of the data to be analyzed.

37.2 Data

The data were obtained from the students records office of a Greek university and their full description is given in Kalamatianou and McClean (2003). The main variable of interest is duration of studies. These are 10313 observations which represent study times of students who entered any department of the university from the academic year 1983-84 until 1992-93.

In the next section we employ kernel smoothing techniques to analyze these data further.

37.3 Estimation

Suppose we have a sample $X_1, X_2, \ldots, X_n$ of i.i.d. survival times censored at the right by i.i.d. random variables $U_1, U_2, \ldots, U_n$, which are independent from the X_i's. Let f be the common probability density and F the distribution function of the X_i's. Also, denote by H the distribution function of the U_i's. Typically the randomly right censored observed data are denoted by the pairs (X_i, Δ_i), $i = 1, 2, \ldots, n$ with $X_i = \min\{X_i,\ U_i\}$ and $\Delta_i = 1_{\{X_i \leq U_i\}}$ where $1_{\{\cdot\}}$ is the indicator random variable of the event $\{\cdot\}$.

An estimate of the unknown pdf f can be defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{H}^*(X_i)} K\left(\frac{x - X_i}{h}\right)$$

where K, called kernel, is a function that integrates to 1, h, called bandwidth, is the amount of smoothing applied to the estimator and $\hat{H}^*$ is an estimate of $1 - H$, typically taken to be the Kaplan-Meier estimator, i.e.

$$\hat{H}^*(x) = \begin{cases} 1, & 0 < x \leq Z_1 \\ \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & Z_{k-1} < x \leq Z_k, k = 2, \ldots, n \\ 0, & Z_n < x \end{cases}$$

with (Z_i, Λ_i) being the ordered (X_i, Δ_i), $i = 1, \ldots, n$. Estimator $\hat{f}(x)$ has been widely discussed in the literature. See Marron and Padgett (1987) for motivation and the references therein for an overview. Practical implementation of $\hat{f}(x)$ requires selection of the kernel K and the bandwidth h. Of the two, of greater importance is selection of the smoothing parameter as this affects the asymptotic properties of the estimate and its visual performance, e.g. Wand and Jones (1995, page 13). All work here uses the Epanechnikov kernel,

$$K(x) = \frac{3}{4\sqrt{5}} \left(1 - \frac{x^2}{5}\right), \quad -\sqrt{5} \leq x \leq \sqrt{5}.$$

Motivation for use of this particular kernel function comes from its optimality properties as those are described in Wand and Jones (1995). Bandwidth selection is typically done by choosing h which minimizes some error criterion. Here we use the least squares cross-validation method, developed for the case of pdf estimation from censored data by Marron and Padgett (1987). We employ the method to minimize the Integrated Squared Error (ISE) for the reasons exhibited in Marron and Padgett (1987). The objective is to choose h which minimizes the Integrated Squared Error (ISE) of $\hat{f}(x)$ given by

$$\mathrm{ISE}(\hat{f}(x)) = \int \hat{f}^2(x)w(x)\,dx - 2\int \hat{f}(x)f(x)w(x)\,dx + \int f(x)\,dx \qquad (37.3.1)$$

where $w(x)$ is a weight function and its purpose is to eliminate endpoint effects. As the third term on the RHS of (37.3.1) is independent of $\hat{f}$ we want to choose h which minimizes the sum of the first two terms. The first of the two terms is known. As about the second, least squares cross validation principles suggest estimating it by

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(X_i)\frac{w(X_i)}{\hat{H}^*(X_i)}1_{\{\Delta_i=1\}}$$

where $\hat{f}_i(x)$ is the 'leave-one-out' version of $\hat{f}$ given by

$$\hat{f}_i(x) = \frac{1}{(n-1)h}\sum_{j\neq i}\frac{1}{\hat{H}^*(X_j)}K\left(\frac{x-X_j}{h}\right)1_{\{\Delta_i=1\}}.$$

Thus we want to choose h which minimizes the cross validation criterion, $\mathrm{CV}(h)$, given by

$$\mathrm{CV}(h) = \int \{\hat{f}(x)\}^2 w(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_i(X_i)\frac{w(X_i)}{\hat{H}^*(X_i)}1_{\{\Delta_i=1\}}.$$

By simple algebra, a more efficient in computation form of $\mathrm{CV}(h)$ is

$$\mathrm{CV}^*(h) = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1_{\{\Delta_i=1\}}}{\hat{H}^*(X_i)\hat{H}^*(X_j)}K_2\left(\frac{X_i-X_j}{h}\right) + \frac{2}{nh}K(0)$$

where

$$K_2(x) = K^*(x) - 2K(x).$$

Minimization of $\mathrm{CV}^*(h)$ typically is done by a grid search for h in the interval $n^{-1/5}\sigma/4 < h < 3n^{-1/5}\sigma/2$ and then extend the interval if the minimum is at the endpoints of either side. After the best point is found a possible improvement would be a quasi-Newton approach.

With h chosen in this manner the ISE of the estimator becomes asymptotically optimal (see theorem 4.1, Marron and Padgett (1987)) in the sense that

$$\frac{\mathrm{ISE}(\hat{f},\hat{h})}{\inf_h \mathrm{ISE}(\hat{f},h)} \to 1 \text{ a.s.}$$

An estimate of the survival function $S(x) = 1 - F(x)$, can be obtained by integrating $\hat{f}(x)$. Let $\hat{S}_n(x) = 1 - \hat{F}(x)$, where $\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(u)\, du$. Estimator $\hat{S}_n(x)$ has been studied by Kulasekera et al (2001). They proved that $\hat{S}_n(x)$ performs better in Mean Squared Error than the Kaplan-Meier estimator.

Then an estimate of the hazard rate function $\lambda(x) = f(x)/(1 - F(x))$ can be obtained by using the estimates $\hat{f}(x)$ and $\hat{F}(x)$ and substituting to $\hat{\lambda}(x)$. This gives $\hat{\lambda}(x) = \hat{f}(x)/(1 - \hat{F}(x))$. Next we apply estimators $\hat{S}_n(x)$ and $\hat{\lambda}(x)$ to the dataset discussed in section 37.2.

37.4 Results

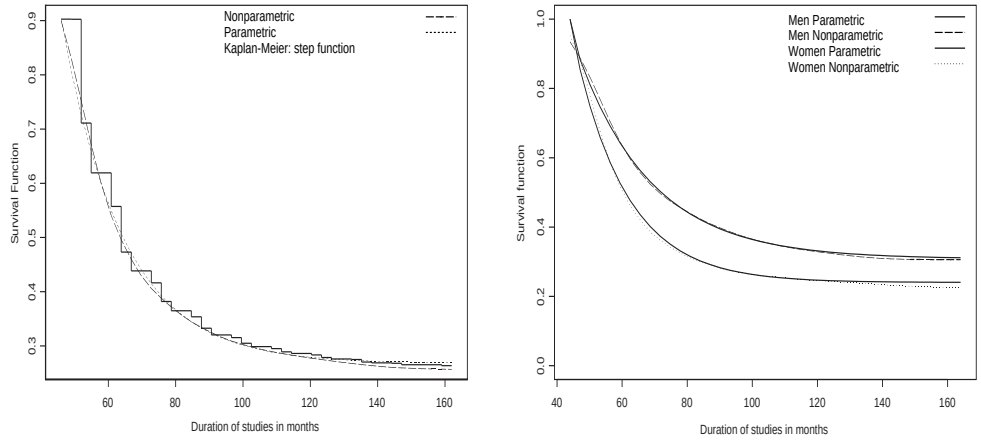
In this section we implement estimators $\hat{S}_n(x)$ and $\hat{\lambda}(x)$ to estimate the survival function and the hazard rate of the total population and the subgroups of men and women. In all cases bandwidth selection was done with the methods described in section 37.3. For survival function estimation we use for comparison the parametric estimates of Kalamatianou and McLean (2003). These are $S_a(x) = 0.26801 + 0.6344e^{-0.05471(x-46)}$ for the total population, $S_w(x) = 2401 + 0.6566e^{-0.06198(x-46)}$ for the women subgroup and $S_m(x) = 0.3081 + 0.6034e^{-0.04398(x-46)}$ for the male subgroup. In all cases $x \geq 46$.

Interpretation of the estimated survival function is as follows: The values of the survival function express the probability that a student will not graduate after time x. Therefore, small values of the estimated survival function at time x mean high probability for someone to graduate while the opposite happens for large values.

In figure 37.1(a) we plot the estimated survival function $\hat{S}_n(x)$ for the total population together with the Kaplan-Meier estimator and estimator $S_a(x)$. We see that the continuous estimates perform quite similarly. Both estimates suggest a similar rate of decrease from the beginning of the interval of estimation all the way through to the end.

In figure 37.1(b) we use $\hat{S}_n(x)$ as well as $S_w(x)$ and $S_m(x)$ to estimate the survival function of the subgroups of women and men. Again both estimators for both populations estimators perform quite similarly. The main conclusion is that there are indeed differences between the sexes in the graduation process and the duration of studies. As we see from the plot women graduate faster than men.

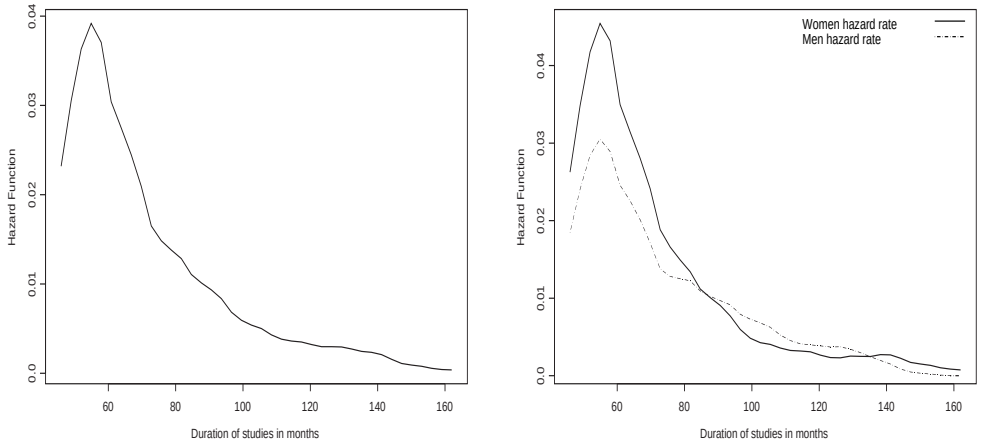
We turn to estimation of the hazard rate function. In general, $\hat{\lambda}(x)$ gives an estimate of the instantaneous probability that a student having not graduated



(a) Kaplan Meier, kernel and parametric estimate

(b) Parametric and nonparametric estimates

Figure 37.1: Parametric and nonparametric estimates of the survival function for the total population and separately for men and women subgroups.



(a) Hazard rate for the total population

(b) Hazard rate for men and women

Figure 37.2: Hazard rate for the total population and separately for men and women subgroups.

at time x will graduate in the time interval $(x, x + \Delta x)$. In figure 37.2(a) we estimate the hazard rate for the total population whereas in figure 37.2(b) we estimate the hazard rate for the subgroups of men and women. All estimates were calculated by using $\hat{\lambda}(x)$. In both figures the pattern is the same for all three hazard estimates, i.e. an initial increase in hazard reaches its peak at the 54th month of study and then the curve decreases. It is apparent from figure 37.2(b) that women graduate faster than men. Note though that this seems to change between the 86th and the 136th month of study, something not indicated by the study of the survival function.

An interesting feature from figure 37.2(a) is that there is a certain number of students that they don't get their degrees even though they have completed 162 months of study. Furthermore figure 37.2(b) suggests that this group consists of both male and female students.

References

1. Kalamatianou, A. and McClean, S. (2003). The perpetual student: Modelling duration of undergraduate studies based on lifetype-type educational data, *Lifetime data analysis*, **9**, 311–330.

2. Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations., *Journal of the American Statistical Association*, **53**, 457–481.

3. Kulasekera, K.B., Williams, C.L., Coffin, M. and Manatuga, A. (2001) Smooth estimation of the reliability function *Lifetime data analysis*, **7**, 415–433.

4. Marron, J.S. and Padgett, W.J. (1987) Asymptotically optimal bandwidth selection for kernel density estimators from randomly right censored samples. *Annals of Statistics*, **15(4)**, 1520–1535.

5. Wand, M. and Jones, M.C. (1995). *Kernel smoothing*, Chapman and Hall, London.

38

Methods for Meta-analysis of Population-based Genetic Association Studies

Pantelis G. Bagos[1] and Georgios K. Nikolopoulos[2]

[1] *Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens, 15701, Greece*
[2] *Hellenic Centre for Diseases Control and Prevention, 3rd September 54, Athens, 10433, Greece*

Abstract: We propose a simple and robust approach for meta-analysis of population-based genetic association studies. The method exploits the binary nature of the data, and treats the genotypes as independent variables in a logistic regression model. We present simple tests for detecting heterogeneity and we describe a random effects extension of the method using multilevel modeling in order to allow for between studies heterogeneity. We derive also simple methods for deciphering the genetic model of inheritance and adjusting for potential confounders. The methodology was applied in three published meta-analyses with very promising results. The proposed approach is flexible and easy to use from the one hand, while at the other hand it covers almost every aspect of a meta-analysis providing overall estimates and avoiding multiple comparisons. We expect that this simple method would be used in the foreseeable future in meta-analyses of gene-disease association studies.

Keywords and phrases: Meta-analysis, genetic epidemiology, random effects, logistic regression

38.1 Introduction

Meta-analysis of genetic association studies is performed in order to investigate the relation of a particular genetic marker and a disease synthesizing the available information in the field. The main difference compared to the well-known epidemiological studies, is the fact that the "exposure" (i.e. the genotype) has more than two levels. In Table 1, we list the data corresponding to a meta-analysis of four studies regarding the association of KIR6.2 E23K polymorphism and Type II diabetes taken from Hani et al., (1998). Traditional methods of meta-analysis, Petiti, (1994), require that the three genotypes should be collapsed into two categories (for instance AA vs AB+BB) or that pair-wise com-

parisons should be made (BB vs AB and so on). The choice of the categories to be combined is important since it depends on the underlying genetic model of inheritance. Nevertheless, even in the best case the risk of performing multiple tests cannot be disregarded.

	Genotype					
	Controls			Cases		
Study	**AA**	**AB**	**BB**	**AA**	**AB**	**BB**
1	38	52	6	72	78	22
2	44	27	11	38	45	17
3	33	34	8	21	26	11
4	45	53	16	53	87	51

Table 1. Data taken from the meta-analysis of Hani *et al.,* (1998) concerning the association of KIR6.2 gene polymorphisms with Type II diabetes (A stands for Glutamic, B stands for Lysine).

Thakkinstian et al., (2005), proposed a methodology of predefined steps using the commonly used approach of meta-analysis with summary data. Minelli et al., (2005b), proposed the so-called (genetic) model-free approach, which does not assume a genetic model a priori but instead deduces it from the data. More specifically, they modelled jointly the ORBB, which is the logOR of BB genotype vs AA, and ë, which is the ratio of logORBB and logORAB (i.e the OR of AB genotype vs BB), an approach that recognizes the fact the two ORs are correlated. Recently, Minelli et al., (2005a) extended their method in a Bayesian framework investigating the use of both prospective and retrospective likelihood and concluded that both methods produce equivalent results. Here, we will present a simple alternative to the aforementioned methods, using the genotypes as independent variables in a logistic regression. The method is easily performed in nearly any statistical software, and with this approach we overcome the problem of multiple comparisons between genotype contrasts as well as the non-normality of summary measures (logORs). Furthermore, there is no need for specialized software in order to fit the more sophisticated models.

38.2 Logistic regression models

38.2.1 Fixed effects logistic regression

Let y_{ij} denote the number of cases, n_{ij} the total number of subjects, and π_{ij} the underlying risk of the j^{th} person in the i^{th} study respectively. Considering allele B as the risk factor, the AA genotype ($r = 1$) is treated as the reference category and we create dummy variables such as $z_{2ij} = 1$ if the genotype is AB

($r = 2$) and $z_{3ij} = 1$ if the genotype is BB ($r = 3$). Using ordinary logistic regression we can perform a meta-analysis stratified by studies as follows:

$$logit(\pi_{ij}) = \alpha_i + \theta_2 z_{2ij} + \theta_3 z_{3ij} \tag{1.1}$$

where, we include dummy variables α_i as indicators of the study-specific fixed-effects. The exponentiated coefficients θ_2 and θ_3 here provide the combined estimate for the odds-ratio associated with each genotype. The model in Equation (1.1) assumes the homogeneity of the genotype effects between studies. Attaching a term for the interaction between the study effect and the genotypes:

$$logit(\pi_{ij}) = \alpha_i + \sum_{c=2}^{r} \theta_c z_{cij} + \sum_{i=2}^{k} \sum_{c=2}^{r} \gamma_{ic} \alpha_i z_{cij} \tag{1.2}$$

corresponds to testing the hypothesis that the effect of each genotype vary significantly between studies. This hypothesis can be tested by applying a multivariate Wald test, where the null hypothesis is:

$$H_0 : \gamma_{ic}, \forall i = 2, 3, \ldots, k; c = 2, 3, \ldots, r.$$

Denoting by $\mathbf{b}$ the vector of the estimated coefficients, by $\mathbf{V}$ the estimated variance-covariance matrix, and by $\mathbf{Rb} = \mathbf{r}$ the vector of the $(c-1)(i-1)$ linear hypotheses, the statistic:

$$W = (\mathbf{RB} - \mathbf{r})'(\mathbf{RVR}^{\star})^{-1}(\mathbf{Rb} - \mathbf{r}) \tag{1.3}$$

will have asymptotically a χ^2 distribution as shown by Judge et al., (1985):

$$W \ \chi^2_{(i-1)(c-1)}. \tag{1.4}$$

This test for the significance of the interaction terms is the analogue of the χ_2 test for heterogeneity (Cochran's Q) used in a summary data method. The Wald test could be used also for testing contrasts between the derived coefficients. Moreover, W could be used for calculating a modified version of the inconsistency index I^2, initially proposed by Higgins et al., (2003):

$$I^2 = \frac{W - (i-1)(c-1)}{W} 100\%. \tag{1.5}$$

38.2.2 Random effects logistic regression

In order to consider an additive component of heterogeneity, and fit a random-effects logistic regression allowing the variability of the genotype effects between studies, we introduce a study-specific random coefficient v_{ci}, representing the deviation of study's true effect (θ_{ri}) from the overall mean effect θ_r, thus:

$$logit(\pi_{ij}) = \alpha_i + (\theta_2 + v_{2i}) z_{2ij} + (\theta_3 + v_{3i}) z_{3ij} \tag{1.6}$$

In the above model, the random terms are distributed normally:

$$\begin{pmatrix} v_{2i} \\ v_{3i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{2i}^2 & \sigma_{23i}^2 \\ \sigma_{32i}^2 & \sigma_{3i}^2 \end{pmatrix} \right) \tag{1.7}$$

with the covariance between the random terms being equal to:

$$Cov(v_{2i}, v_{3i}) = \sigma_{23i} = \sigma_{32i} = \rho\sigma_{2i}\sigma_{3i} \tag{1.8}$$

where ρ is the correlation of the two random terms. Variations on this model include, the incorporation of random study-specific effects, or a common variance for the random coefficients. All these approaches need different coding for the dummy variables, and are not explored further. The reader is referred to the works of Turner et al., (2000); Higgins et al., (2001). The logistic regression model finally, could be used to infer the genetic model. For instance, in case when both θ_2 and θ_3 are significantly greater than zero then we can test the null hypothesis:

$$H_0 : \theta_2 = \theta_3 \tag{1.9}$$

using a Wald test as the one described above with a chi-square distribution with 2 d.f. A non-significant p-value (> 0.05) indicates the equality of the logORs, suggesting a dominant mode of inheritance.

38.3 Application on published meta-analyses

We present here an application of the proposed methodology in three already published meta-analyses and we compare the obtained results with those of the other methods (Table 2). We evaluate the traditional methods using summary data, under both fixed effects (FE) and random effects (RE) assumptions, the genetic model-free approach with random-effects with λ bounded ($0 < \lambda < 1$) and unbounded, as well as the Bayesian approach under prospective and retrospective likelihood.

In nearly all the cases the methods yield comparable results. In meta-analyses with smaller number of included studies, the Bayesian methods produce wider confidence intervals reflecting the uncertainty of the priors used. In the case of the meta-analysis of Hani et al., (1998) concerning the association of KIR6.2 gene polymorphisms and Type II diabetes, which consists of only 4 studies, the Bayesian methods conclude that there is no overall association. In the same dataset, the traditional methods based on summary data, clearly overestimate the presumed risk.

The model-free approach, with bounded λ tends to produce always confidence intervals not including one. A major difference with the logistic regression approach, appears in the meta-analysis of Kato et al., (1999) regarding the

relationship between the M235T AGT polymorphism and Essential hypertension. Here, the random effects logistic regression model provided a significantly higher estimated risk associated with the BB (i.e. TT) genotype compared to all the other methods. This is clearly a consequence of the non-normality of the distribution of the particular logOR (p-value according to the Shapiro-Wilk test equals to 0.00709), and strongly supports the usefulness of the proposed approach. The genetic model for this study is clearly a recessive one, which comes also in agreement with the results of the two different approaches of Minelli and coworkers. However in the model-free approach with bounded ë as well as in the Bayesian methods, the confidence interval for logORAB does not include one, even though the recessive model corresponds to a non-significant estimate, complicating further the interpretation of the results.

Meta-analysis	**Studies**	**Method**	$\mathbf{OR}_{BB}$**(95% CI)**	$\mathbf{OR}_{AB}$**(95% CI)**
Hani *et al.,* (1998)	4	Summary methods (FE)	2.24 (1.81, 2.67)	1.29 (0.99, 1.58)
		Summary methods (RE)	2.23 (1.79, 2.68)	1.31 (0.84, 1.79)
		Model-free approach (RE, unbounded λ)	2.14 (1.39, 3.29)	1.21 (0.90, 1.63)
		Model-free approach (RE, bounded λ)	2.14 (1.43, 3.29)	1.21 (1.08, 1.63)
		logistic regression (FE)	2.15 (1.40, 3.30)	1.21 (0.91, 1.63)
		logistic regression (RE)	2.15 (1.40, 3.30)	1.22 (0.91, 1.63)
		Bayesian method (prospective)	2.01 (0.97, 4.09)	1.16 (0.99, 1.77)
		Bayesian method (retrospective)	2.03 (0.96, 3.96)	1.16 (0.99, 1.80)
Kato *et al.,* (1999)	7	Summary methods (FE)	2.24 (1.81, 2.67)	1.29 (0.99, 1.58)
		Summary methods (RE)	1.78 (1.10, 2.89)	1.10 (0.72, 1.66)
		Model-free approach (RE, unbounded λ)	1.64 (0.99, 2.72)	1.00 (0.66, 1.53)
		Model-free approach (RE, bounded λ)	1.64 (1.15, 3.05)	1.00 (1.00, 1.62)
		logistic regression (FE)	1.67 (1.13, 2.46)	1.16 (0.78, 1.74)
		logistic regression (RE)	1.95 (1.24, 3.08)	0.98 (0.60, 1.57)
		Bayesian method (prospective)	1.81 (1.05, 3.66)	1.08 (1.00, 1.74)
		Bayesian method (retrospective)	1.83 (1.06, 3.60)	1.09 (1.00, 1.71)
Wheeler *et al.,* (2004)	19	Summary methods (FE)	2.24 (1.81, 2.67)	1.29 (0.99, 1.58)
		Summary methods (RE)	1.14 (0.99, 1.32)	1.10 (0.99, 1.23)
		Model-free approach (RE, unbounded λ)	1.17 (1.04, 1.33)	1.08 (1.00, 1.17)
		Model-free approach (RE, bounded λ)	1.17 (1.04, 1.33)	1.08 (1.01, 1.17)
		logistic regression (FE)	1.15 (1.02, 1.30)	1.09 (1.01, 1.17)
		logistic regression (RE)	1.14 (1.00, 1.29)	1.09 (1.01, 1.19)
		Bayesian method (prospective)	1.15 (1.01, 1.33)	1.08 (1.00, 1.21)
		Bayesian method (retrospective)	1.15 (1.02, 1.34)	1.08 (1.00, 1.21)

Table 2. The results obtained in the three published meta-analyses. FE: fixed effects; RE: random effects. For explanation of the methods see the text.

In the meta-analysis concerning the association of Paraoxonase (PON1) Q192R polymorphism with Myocardial Infarction by Wheeler et al., (2004), the estimates of the ORs and the confidence intervals are nearly identical in all methods. However, the model-free approach indicates clearly a co-dominant

model of inheritance ($\lambda = 0.53$), contradicting our results, which suggest an intermediate risk associated with the two genotypes. Once again the normality assumptions are crucial for this decision since logORAB is not distributed normally ($p = 0.00035$). Indeed, analysing the same data under the Bayesian models, we see that similar conclusions to ours could be drawn ($\lambda = 0.62$), a result that further strengthen the belief for the validity of the logistic regression approach.

38.4 Conclusions

The methodology introduced here can be easily applied in the meta-analysis of population-based genetic association studies. We presented both fixed and random effects models using the genotypes as independent variables in a logistic regression. The methods are quite familiar to epidemiologists and the results can be interpreted without serious difficulties. Furthermore, we provided tests for assessing heterogeneity, as well as statistical procedures to discover a possible model of inheritance. Compared to the widely used approach in the literature, the proposed methodology is far more robust and permits making overall inferences from the meta-analysis. The methodology proposed by Thakkinstian et al., (2005), is a more careful and detailed version of the widely used approach; however it has many limitations such as the need of multiple comparisons, the lack of an overall test for homogeneity, or of a test for the genetic model and the inability of incorporating covariates simultaneously. Compared to the more sophisticated model-free approaches of Minelli and coworkers, the logistic regression approach is flexibly and easily implemented. Moreover, it allows the incorporation of covariates as well as implements formal overall tests for heterogeneity, which were not addressed by this method. Additionally, the model-free approach demands further modifications in cases where there are more than three genotypes. Finally, all the above-mentioned approaches are based on summary data methods. The logistic regression approach on the other hand, uses directly the binary structure of the data, hence it is expected to perform better in cases where the normality assumption of the logORs is violated.

References

1. Hani, E.H., Boutin, P., Durand, E., Inoue, H., Permutt, M.A., Velho, G. and Froguel, P. (1998) Missense mutations in the pancreatic islet beta cell inwardly rectifying K+ channel gene (KIR6.2/BIR): a meta-analysis suggests a role in the polygenic basis of Type II diabetes mellitus in Caucasians. Diabetologia, 41, 1511-1515.

2. Higgins, J.P., Thompson, S.G., Deeks, J.J. and Altman, D.G. (2003) Measuring inconsistency in meta-analyses. Bmj, 327, 557-560.
3. Higgins, J.P., Whitehead, A., Turner, R.M., Omar, R.Z. and Thompson, S.G. (2001) Meta-analysis of continuous outcome data from individual patients. Stat Med, 20, 2219-2241.
4. Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T.-C. (1985) The Theory and practice of Econometircs. John Wiley & Sons.
5. Kato, N., Sugiyama, T., Morita, H., Kurihara, H., Yamori, Y. and Yazaki, Y. (1999) Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies. J Hypertens, 17, 757-763.
6. Minelli, C., Thompson, J.R., Abrams, K.R. and Lambert, P.C. (2005a) Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. Stat Med, 24, 3845-3861.
7. Minelli, C., Thompson, J.R., Abrams, K.R., Thakkinstian, A. and Attia, J. (2005b) The choice of a genetic model in the meta-analysis of molecular association studies. Int J Epidemiol, 34, 1319-1328.
8. Petiti, D.B. (1994) Meta-analysis Decision Analysis and Cost-Effectiveness Analysis. Oxford University Press.
9. Thakkinstian, A., McElduff, P., D'Este, C., Duffy, D. and Attia, J. (2005) A method for meta-analysis of molecular association studies. Stat Med, 24, 1291-1306.
10. Turner, R.M., Omar, R.Z., Yang, M., Goldstein, H. and Thompson, S.G. (2000) A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med, 19, 3417-3432.
11. Wheeler, J.G., Keavney, B.D., Watkins, H., Collins, R. and Danesh, J. (2004) Four paraoxonase gene polymorphisms in 11212 cases of coronary heart disease and 12786 controls: meta-analysis of 43 studies. Lancet, 363, 689-695.

39

Bayesian Models for Safety Design to Prevent Foreign Body Injuries in Children

P. Berchialla[1], S. Snidero[2], A. Stancu[2], C. Scarinzi[2], R. Corradetti[2], D. Gregori[1]

[1]*Department of Public Health and Microbiology, University of Torino*
[2]*Department of Statistics and Applied Mathematics, University of Torino*

Abstract: The entry of a small item into the upper aero-digestive ways is one of the leading causes of injuries in children up to 14 years old. The aim of this paper is to show how the Bayesian models along with Markov chain Monte Carlo techniques can be used to formulate a model for use in a quantitative risk assessment. Inference, in the light of evidence, can be made on all domain variables making it possible to sample from the distributions of variables of interest such as volume or shape of objects which caused injuries. Results show how the knowledge of such distribution can be helpful in implementing a safety design of the products.

Keywords and phrases: Quantitative risk assessment; Bayesian models; Foreign body injury

39.1 Introduction

The accidents due to the inhalation, ingestion and aspiration of foreign bodies are still one of the leading causes of injury and death in children [1]. Some of the most commons objects which cause foreign body injuries include balls, marbles and beads, nuts and seeds, fish bones, pebbles, stones and small part of toys.

Many studies have been carried out to characterize the types, shapes, and sizes of objects causing injuries. In [2], a statistical analysis was performed on the features of objects causing choking in children. In [3], a computerised models of the airways and oral cavities of children of various ages was developed in order to asses the hazards of toys and small parts. In [4] the Quantitative Risk Analysis methodology was adopted to evaluate the potential risk associated with any given consumer product. A risk equation was set up to determine the

effect of object characteristics on the risk of injury and predict the probability of injury by multiplying the hazard associated with an object by the exposure of the child to the object. Finally Monte Carlo simulations were utilized to generate estimates of product-related risk.

In general, probabilistic methods enable the characterization of uncertainty associated with the dimensions and the shape of the objects involved in the injuries. Using these methods the safety design of products can be assessed on a quantitative basis, furthermore allowing the evaluation of the risk associated to the characteristic of an object such as its volume and shape. The aim of this paper is to give a quantitative risk assessment for the identification of the features of such products which don't provide anyone who might come into contact with a level of safety.

One of the most challenging aspects of applying any probabilistic methodology to this problem is the determination of the appropriate distribution of the actual features of the products. Nevertheless, one of the benefits of utilizing a probabilistic approach is that Bayesian statistical tools can be used to update probability distributions as soon as new data become available. With regard to the substantive issue motivating the paper, available data are usually coming from official discharge records, although new data collection strategies are being implemented at European level. Following presentation of the model built up, results will be discussed.

39.2 Materials and Methods

39.2.1 Data

The European Survey on Foreign Bodies Injuries Study collected data on foreign body injuries from 19 European Hospitals (Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Finland, Germany, Greece, Italy, Poland, Romania, Slovakia, Slovenia, Spain, Sweden, Swiss, Turkey and United Kingdom). Data on 2103 injuries occurred in the years 2000-2002 were gathered according to the ICD931 to ICD935. Objects were characterised by size, shape and consistency [1]. According to their shape they were assigned to one of the following four categories: Spherical - e.g. ball, pebble; Three-dimensional (3D) - e.g. pen cap; Two-dimensional (2D) - e.g. sheet, cellophane; 2D-circle - e.g. coin. With regard to the size, when the dimensions (expressed in millimetres) of the object were reported, the volume was calculated accordingly to the shape of the objects itself, e.g. for three-dimensional objects the volume of an ellipsis was calculated by the length of the axis, for spherical objects the volume of a sphere was calculated by the diameter reported and finally for two-dimensional circle objects the volume was approximated by that one of a cylinder. Such volume measures represent how much space the smallest geometrical figure containing

the irregular-shaped foreign body takes up. In addition to the volume, the ratio between long axis and short axis was calculated in order to make the ellipticity of a measured foreign body available.

39.2.2 Bayesian model

A multivariate Bayesian model was used to model volume and ellipticity. Since volume and ellipticity are positive values with an empirical distribution which is skewed to the right, their log transformed was assumed to be normal. A Wishart(R,ρ) prior was specified for the population matrix of the parameters. To represent a vague prior knowledge, we chose the degrees of freedom ρ for this distribution to be as small as possible—i.e. 2, the rank of Ω.

The model was expressed in the equation form as:

$$X|\theta, \tau \sim Normal(\mu, \tau^{-1}) \tag{39.2.1}$$

$$\theta \sim Normal(\mu, \Omega^{-1}) \tag{39.2.2}$$

$$\tau \sim Wishart(R, \rho) \tag{39.2.3}$$

Missing values were generated from their full conditional distributions through the use of the Gibbs algorithm. To ensure stability of the results, the Gibbs sampler was run for 61,000 updates with the first 10,000 discarded as burn–in. The benefit of this approach is in its generality. In fact it was assumed that only the data come from a log–normal distribution with a mean for each characteristics and an overall variance.

Bayesian statistics provides a very plain approach to a "learning from experience"process which allows new data to be used in order to revise baseline probability distributions. Suppose that a surveillance system observes a new data point. The model can be updated through the classical Bayesian approach:

$$P(data_{new}|data) = \int P(data_{new}|\theta, data)d\theta = \int P(data_{new}|\theta)P(\theta)d\theta \tag{39.2.4}$$

39.3 Results

In figure 39.1 is shown the bivariate density of volume and ellipticity sampled from the population of foreign body which caused injuries in children. In figure 39.2 is shown the marginal density of volume and ellipticity. In table 39.1 summary statistics of volume and ellipticity parameters are shown.

Any sample from the joint probability of volume and ellipticity represents a collection of volume and ellipticity data from the population of products which caused injuries. From the posterior density it is straightforward to detect the item characteristics which pose a major threat to the children health.

Sampling from the posterior distribution of volume and ellipticity, the highest probability is associated to a volume lesser than 631 millimeters^3 and an ellipticity between 0.97 and 1.085—i.e. spherical shaped objects with a diameter of about 8.5 millimeters.

39.4 Conclusion

The evolution of the food industry has contributed to the introduction of new forms of presentation and packaging, leading to the combination of edible and inedible components, such as toys, which may pose a hazard to consumer safety. In this paper, it has been shown how the Bayesian modeling can be used to formulate a model for use in a quantitative risk assessment aimed to implement a safety design which is determined by whether a product provides anyone who might come into contact with it a level of safety. In order to do so,it could be important to model item characteristics such as volume and ellipticity by severity of injury. In fact joint probabilities of item features and severity can better characterize the overall risk than marginal or conditional probabilities.

References

1. Gregori D., M. B., Snidero S., Corradetti R., Passali D. (2004). Foreign bodies in aero-digestive tract: a meta-analysis 1973-2003. submitted to European Journal of Pediatric Otorhinolaryngology.

2. Rimell, F. L., Thome Jr, A., Stool, S., Reilly, J. S., Rider, G., Stool, D. and Wilson, C. L. Characteristics of objects that cause choking in children. Journal of American Medical Association (1996). 274(22):1763-1766.

3. Stool, D., Rider, G. and Welling, J. R. Human factors project: development of computer models of anatomy as an aid to risk management. International Journal of Pediatrics and Otorhinolaryngology (1998). 43(3):217-227.

4. Rider, G., Milkovich, S., Stool, D., Wiseman, T., Doran, C. and Chen, X. Quantitative risk analysis. Injury Control and Safety Promotion (2000). 7(2):115-133.

5. Spiegelhalter, D. J., Thomas, A., Best, N., Lunn, D. WinBUGS Version 1.4 User Manual. MRC Biostatistic Unit, Cambridge, UK. 2003.

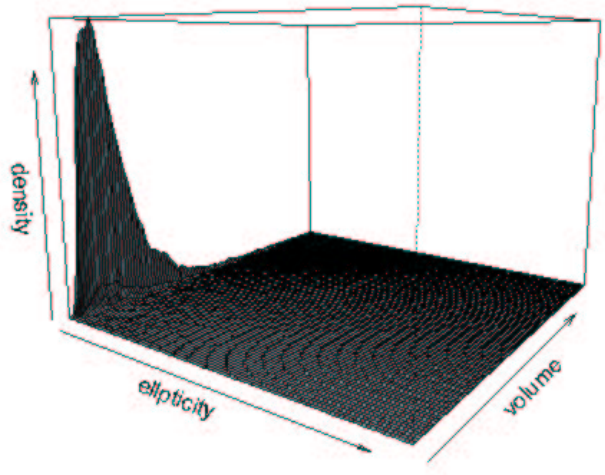


Figure 39.1: Bivariate density of volume and ellipticity

Table 39.1: Volume and ellipticity parameters estimates

*Parameters	Mean	sd	2.5%	Median	97.5%
volume	106.27	1.08	92.76	106.27	54.60
ellipticity	1.33	2.05	1.28	1.34	1.41

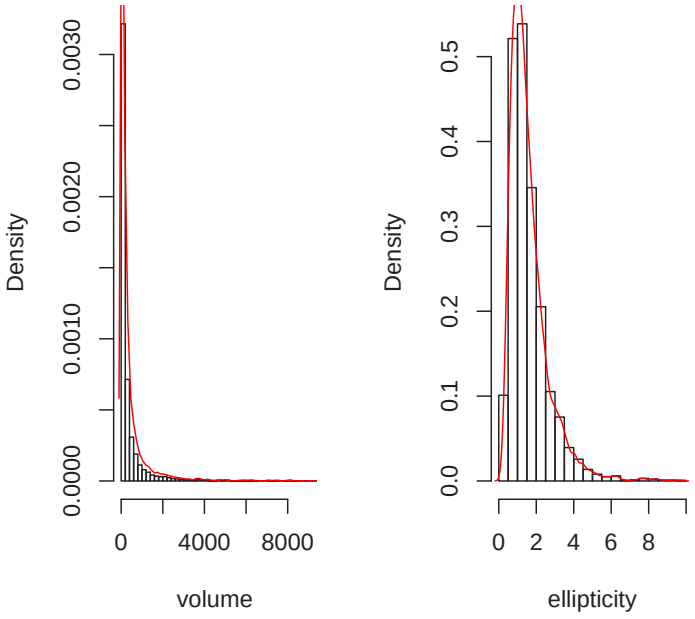


Figure 39.2: Marginal density of volume and ellipticity

40

Progressive Type-II Censoring And Transition Kernels

Eric Beutner

Institute of Statistics, RWTH Aachen University, Germany

Abstract: Progressively type-II censored order statistics are widely used in reliability theory and biometrics. Using transition kernels we provide a new definition of progressively type-II censored order statistics which allows an easy derivation of known properties of these random variables. It is shown that, for an absolutely continuous distribution function, our definition of progressively type-II censored order statistics leads to a joint density which coincides with the joint density of progressively type-II censored order statistics derived by Viveros and Balakrishnan (1994).
Moreover, we extend the model by allowing not only to remove units from the experiment but also to add additional units to the experiment.

Keywords and phrases: Progressive type-II censoring, Modified progressive type-II censoring, Transition kernels, Markov chains

40.1 Introduction

Suppose that a total of n units is placed on a life-test. Let the iid random variables $X_1, \ldots, X_n$ describe the failure times of these units. We assume that the distribution function F of these random variables is absolutely continuous and denote its density by f. The well known progressive type-II censoring scheme works as follows: At the time of the first failure a number of R_1 randomly chosen units of the remaining $n-1$ units are removed from the experiment. At the time of the next failure R_2 randomly chosen units of the remaining $n-2-R_1$ are censored, and so on. Finally, at the time of the mth failure all remaining units are removed. So, due to censoring one only observes m failure times which will de denoted by $X_{1:m:n}, ..., X_{m:m:n}$. From their description of progressively type-II censored order statistics under the censoring scheme

$(R_1, \ldots, R_m)$, $R_i \in \mathbf{N}$, $i = 1, \ldots, m$, Viveros and Balakrishnan (1994) (see also Aggarwala and Balakrishnan (2000, p. 7)) derive the following joint density:

$$f_{X_{1:m:n},X_{2:m:n},\ldots,X_{m:m:n}}(x_1, x_2, \ldots, x_m) = c\prod_{i=1}^{m} f(x_i)(1 - F(x_i))^{R_i}, \qquad (40.1.1)$$

for $x_1 < x_2 < \ldots < x_m$ where $c = n(n - R_1 - 1) \cdot \ldots \cdot (n - \sum_{i=1}^{m-1} R_i - m + 1)$. In the following, to simplify the notation we set $\gamma_j = n - \sum_{i=1}^{j-1} R_i - (j - 1)$, $j = 2, \ldots, m$.

40.2 Progressive type-II censoring via transition kernels

The idea underlying our definition of progressively type-II censored order statistics $X_{1:m:n}, \ldots, X_{m:m:n}$ is the following: After the $(j-1)$th failure at time t_{j-1} and after removing $\sum_{i=1}^{j-1} R_i$ items from the experiment we still have γ_j items at work. These items are independent and have distribution function

$$F_j(t) = \frac{F(t) - F(t_{j-1})}{1 - F(t_{j-1})}, \;\; t \geq t_{j-1}. \qquad (40.2.2)$$

The distribution function of the minimum X_{min} of a sample of $n - (j-1) - \sum_{i=1}^{j-1} R_i$ units having distribution function (40.2.2) is

$$\begin{aligned} P(X_{min} \leq t) &= 1 - \left(1 - \frac{F(t) - F(t_{j-1})}{1 - F(t_{j-1})}\right)^{\gamma_j} \\ &= 1 - \left(\frac{1 - F(t)}{1 - F(t_{j-1})}\right)^{\gamma_j}, \;\; t \geq t_{j-1}. \end{aligned} \qquad (40.2.3)$$

Now, for $j = 2, \ldots, m$, define transition kernels p_j on $(\mathbf{R}_+, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-Algebra on $\mathbf{R}_+$ by

$$p_j(s, B) = \int_B \frac{\gamma_j}{(1 - F(s))^{\gamma_j}} f(t)(1 - F(t))^{\gamma_j - 1} I_{[s,\infty)}(t)dt \qquad (40.2.4)$$

and let the initial distribution p be given by

$$p(B) = \int_B nf(x)(1 - F(x))^{n-1}dx. \qquad (40.2.5)$$

Notice that, for $j = 2, \ldots, m$, $p_j(s, \cdot)$ is a probability measure on $\mathcal{B}$ for every $s \in \mathbf{R}_+$, and $p_j(\cdot, B)$ is $\mathcal{B}$-measurable for every $B \in \mathcal{B}$.

Theorem 40.2.1 *There exist a probability space $(\Omega, \mathcal{F}, P)$ and a Markov chain $X_{1:m:n}, \ldots, X_{m:m:n}$ defined on that space such that*

$$P(X_{1:m:n} \leq t) = p([0,t]) \tag{40.2.6}$$

and

$$\begin{aligned} &P(X_{j:m:n} \leq t | X_{1:m:n} = t_1, \ldots, X_{j-1:m:n} = t_{j-1}) \\ &\quad = P(X_{j:m:n} \leq t | X_{j-1:m:n} = t_{j-1}) \\ &\quad = p_j(t_{j-1}, [t_{j-1}, t]) \end{aligned} \tag{40.2.7}$$

for $j = 2, \ldots, m$

Proof. See, e.g., Iosifescu and Tautu (1973, p. 121). ■

Having established the existence of random variables fulfilling (40.2.6) and (40.2.7) we define

Definition 40.2.1 *Random variables $X_{1:m:n}, \ldots, X_{m:m:n}$ with the properties (40.2.6) and (40.2.7) are called progressively type-II censored order statistics (based on F).*

Remark 40.2.1 *For example, it is immediate from our definition that the marginal distribution of $X_{j:m:n}$ is independent of $R_j, \ldots, R_m$ and that $X_{1:m:n}, \ldots, X_{j:m:n}$ are progressively type-II censored order statistics with censoring scheme $(R_1, \ldots, R_{j-1}, n - j - \sum_{i=1}^{j-1} R_i)$.*

The following remark explains how our definition of progressively type-II censored order statistics can be used to define generalized progressively type-II censored order statistics ${}_rX_{r+1:m:n}, \ldots, {}_rX_{m:m:n}$ (cf. Aggarwala and Balakrishnan (2000, p. 9)).

Remark 40.2.2 *Define the initial distribution ${}_rp$ by*

$${}_rp(B) = \int_B \frac{n!}{r!(n-r-r)!} f(x) F(x)^r (1 - F(x))^{n-r-1} dx$$

and, for $j = r+2, \ldots, m$, the transition kernels on $(\mathbf{R}_+, \mathcal{B})$ by

$${}_rp_j(s,B) = \int_B \frac{n - (j-1) - \sum_{i=r+1}^{j-1} R_i}{(1 - F(s))^{n-(j-1)-\sum_{i=r+1}^{j-1} R_i}} f(t) (1-F(t))^{n-(j-1)-\sum_{i=r+1}^{j-1} R_i - 1} I_{[s,\infty)}(t) dt.$$

Theorem 40.2.1 can then be used to establish the existence of a probability space $(\Omega, \mathcal{F}, P)$ and a Markov chain ${}_rX_{r+1:m:n}, \ldots, {}_rX_{m:m:n}$ such that (40.2.6) and (40.2.7) hold with p and p_j replaced by ${}_rp$ and ${}_rp_j$, respectively.

The next theorem shows that our definition of progressively type-II censored order statistics leads to the same joint density as derived by Viveros and Balakrishnan (1994).

Theorem 40.2.2 *The joint density of the random variables of Definition 40.2.1 coincide with (40.1.1).*

Proof. The joint distribution function of the random variables of Definition 40.2.1 is given by (cf. Iosifescu and Tautu (1973, p. 121)

$$\begin{aligned} P(X_{1:m:m} \leq t_1, \ldots, X_{m:m:n} \leq t_m) = \int_0^{t_1} \int_{t_1}^{t_2} \ldots \int_{t_{m-1}}^{t_m} n(1-F(x_1))^{n-1} f(x_1) \\ \cdot \tfrac{\gamma_2}{(1-F(t_1))^{\gamma_2}} f(x_2)(1-F(x_2))^{\gamma_2 - 1} \\ \cdot \ldots \cdot \tfrac{\gamma_m}{(1-F(t_{m-1}))^{\gamma_m}} f(x_m)(1-F(x_m))^{\gamma_m - 1} dx_m \ldots dx_1 \end{aligned}$$

Differentiation with respect to $t_1, \ldots, t_m$ leads to (40.1.1) ■

40.3 Modified progressively type-II censored order statistics

In this section we extend the model by assuming the following situation: Suppose at the time t_{j-1} of the $(j-1)$th failure we can either remove R_{j-1} units from the experiment or add some units (corresponding to $R_{j-1} < 0$) of age t_{j-1} and with the same distribution function F. In that case, the censoring scheme is given by $(R_1, \ldots, R_m)$ where now $R_i \in \mathbf{Z}$ for $i = 1, \ldots, m-1$. If we assume that the added units are independent, and independent of $(X_1, \ldots, X_n)$, the next observed failure has the same distribution as the minimum of a sample of $n - (j-1) - \sum_{i=1}^{j-1} R_i$ units having distribution function (40.2.2). We let again $\gamma_j = n - (j-1) - \sum_{i=1}^{j-1} R_i$, $j = 2, \ldots, m$. Now, define the initial distribution ${}_m p$ by (40.2.5) and the transition kernels ${}_m p_j$ by (40.2.4) where R_i, $1 \leq i \leq m-1$, may now be negative. We use again Theorem 40.2.1 to establish the existence of a probability space $(\Omega, \mathcal{F}, P)$ and random variables ${}_m X_{1:m:n}, \ldots, {}_m X_{m:m:n}$ such that (40.2.6) and (40.2.7) hold with p and p_j replaced by ${}_m p$ and ${}_m p_j$, respectively.

Definition 40.3.1 *Random variables ${}_m X_{1:m:n}, \ldots, {}_m X_{m:m:n}$ fulfilling (40.2.6) and (40.2.7) with p and p_j replaced by ${}_m p$ and ${}_m p_j$, respectively, are called modified progressively type-II censored order statistics (based on F).*

We show next that well known properties of progressively type-II censored order statistics also hold for modified progressively type-II censored order statistics.

Lemma 40.3.1 *Let $F(x) = 1 - e^{-x}$. Then the random variables*

$$Z_1 = n \cdot ({}_mX_{1:m:n}) \tag{40.3.8}$$

and

$$Z_j = \gamma_j \cdot ({}_mX_{j:m:n} - {}_mX_{j-1:m:n}) \;\; j = 2, \ldots, m \tag{40.3.9}$$

are independent and identically distributed as standard exponential.

PROOF. It follows from our definition of modified progressively type-II censored order statistics and the proof of Theorem 40.2.2 that their joint density $f_{{}_mX_{1:m:n},\ldots,{}_mX_{m:m:n}}$ is given by

$$\begin{aligned} f_{X_{1:m:n},X_{2:m:n},\ldots,X_{m:m:n}}(x_1, x_2, \ldots, x_m) &= c \prod_{i=1}^{m} f(x_i)(1 - F(x_i))^{R_i} \quad (40.3.10) \\ &= c \exp\left(-nx_1 - \textstyle\sum_{i=2}^{m} \gamma_i (x_i - x_{i-1})\right), \end{aligned}$$

for $x_1 < x_2 < \ldots < x_m$ where $c = n \cdot \gamma_1 \cdot \ldots \gamma_m$ and $R_i \in \mathbf{Z}$, $1 \leq i \leq m-1$. The result now follows from the definition of Z_j, $1 \leq j \leq m$, and density transformation. ■

Remark 40.3.1 *From the preceding Lemma we immediately obtain a simulation algorithm for modified progressively type-II censored order statistics from an arbitrary distribution function F. Simulate m standard exponential random variables, calculate simulated modified progressively type-II censored order statistics from a standard exponential distribution according to (40.3.8) and (40.3.9) and set ${}_mY_{1:m:n} = F^{-1}(1 - exp({}_mX_{1:m:n})), \ldots, {}_mY_{m:m:n} = F^{-1}(1 - exp({}_mX_{m:m:n}))$.*

If our modified progressively type-II censored order statistics ${}_mX_{1:m:n}, \ldots, {}_mX_{m:m:n}$ are based on the Uniform$(0,1)$ distribution we have the following result

Lemma 40.3.2 *The random variables*

$$V_j = \frac{1 - {}_mX_{m-j+1:m:n}}{1 - {}_mX_{m-j:m:n}}, \;\; j = 1, ..., m, {}_mX_{0:m:n} = 0$$

are independent and V_j is Beta$(n - (m-j) - \sum_{k=1}^{m-j} R_k, 1)$ distributed, $j = 1, \ldots, m$.

PROOF. By analogy with the case of progressively type-II censored order statistics. ■

Remark 40.3.2 *In the definition of the transition kernels there is no need to confine ourselves to natural exponents. We can choose arbitrary positive numbers* $(\alpha_1, ..., \alpha_m)$*. The same is true for the exponent of the initial distribution. Comparing the joint density (40.3.10) of modified progressively type-II censored order statistics with the joint density of generalized order statistics (cf. Kamps (1995)) we see that they are contained in this model in the distribution theoretical sense. Hence, Lemma 40.3.1 may also be obtained from Theorem 3.10 of Kamps (1995).*

References

1. Aggarwala, R., and Balakrishnan, N. (2000). *Progressive Censoring. Theory, Methods, and Applications*, Birkhuser, Boston.

2. Iosifescu, M., and Tautu, P. (1973). *Stochastic processes and applications in biology and medicine I*, Springer, Berlin.

3. Kamps, U. (1995). A concept of generalized order statistics, *Journal of Statistical Planning and Inference*, **48**, 1–23.

4. Viveros, R., and Balakrishnan, N. (1994). Interval Estimation of Parameters of Life From Progressively Censored Data, *Technometrics,* **36**, *84–91.*

41

A Markov Model for Disease Prevalences Including Population Development

Karl-Ernst Erich Biebler and Bernd Paul Jäger

Institute of Biometry and Medical Informatics
Ernst-Moritz-Arndt-University Greifswald
Germany

Abstract: Disease prevalence development shall be predicted. For this purpose, a stochastic processes was applied.
The database is the register of all diabetics of the earlier German Democratic Republic. The register concerns a population of about 17 million people and the period of 1960 to 1989 and covers information on nearly 510 million person-years.
The modelling with the help of Markov chain theory proved to be the best method. For the model parameters, maximum-likelihood-estimators could be applied. It is proved, regarding the properties of the Markov chain model and the data basis, that the observation time period is sufficient, and necessary, to obtain the desired predictions for the development of diabetes prevalences in the population observed.

Keywords and phrases: Disease prevalence prediction, markov chain epidemiological model

41.1 Introduction

One of the advantages of the former Democratic Republic's (GDR) centrally organized public health system was that it allowed for the possibility of establishing a state wide diabetes register over a long period of time.
This unique data collection was held at the former Central Institute for Diabetes in Karlsburg, near Greifswald, Germany. The register covers the period from 1960 to 1989, concerns the former GDR population (about 17 million people) and represents approximately 98% (Michaelis and Jutzi (1991)) of the inflicted persons. It covers information on nearly 510 million person-years.
The prediction of the prevalence development of the diabetes on the base of this register is the central question.

Different regression models were brought into line with the data. The results were not satisfactory.
The essential part of our work lies in the application of the Markov chain theory on the described task. Markov chains are used by MÜLLER for example, for the treatment of similar problems.
Fortunately, four circumstances meet in our situation: There exists a unique database. The Markov chain model fits the data well and, from a mathematical point of view, has good properties. Consequently, the searched parameters can be calculated by the "best" method with maximum precision.
The analysis performed on the Karlsburg diabetes register produced the following main results: The course of time of the prevalences is well described using Markov chains, and the stationary distribution of the stochastic process can be used for the prevalence prediction. It is proved, regarding the mathematical model and the databases, that the observation time period is sufficient, and necessary, to obtain the desired predictions for the development of prevalences.

41.2 Methods

The data collected from 1960 to 1989 at the former Central Institute of diabetes at Karlsburg; Germany, provide the basis of our studies on prevalence of diabetes mellitus in the population of the former GDR.
The data from the total time period were recorded annually by the administrative districts of the country and handed in to the Central Institute in Karlsburg in form of a complete census. The annual reports were divided into the numbers of diabetics, new cases of diabetes, and deceased diabetics according to age group and gender. In addition, it was differentiated between insuline dependent diabetes mellitus (IDDM) and non insuline dependent diabetes mellitus (NIDDM). The administrative incidence is the registration of accessible cases. This is distinguished from the true incidence. Insulin dependence and the centralist structure of the GDR allowed no difference between administrative and true IDDM-incidence. Caution is required with the data recorded as NIDDM. They should be relatively close to the true values, although the variably strong deviations from the true incidence in different years must be considered, since during the time period reported in the GDR screening tests for diabetes mellitus were conducted several times. These are also reflected in the present data material. As a whole MICHAELIS and JUTZI (1991) assumes the degree of registration reliability to be ca. 98 percent.
In these studies, the data available from 1960 to 1989 are separated according to gender (F, M) and type of disease treatment (IDDM or T1, NIDDM or T2). Within these four groups, 3 age classes young (Y, 0-19-yr-olds), middle (M, 20-39-yr-olds) and old (O, over 39-yr-olds) are differentiated. The division chosen was recommended by experienced clinicians.

It was our goal to describe the prevalence of diabetes mellitus over the course of time and to predict, from these studies, disease development.

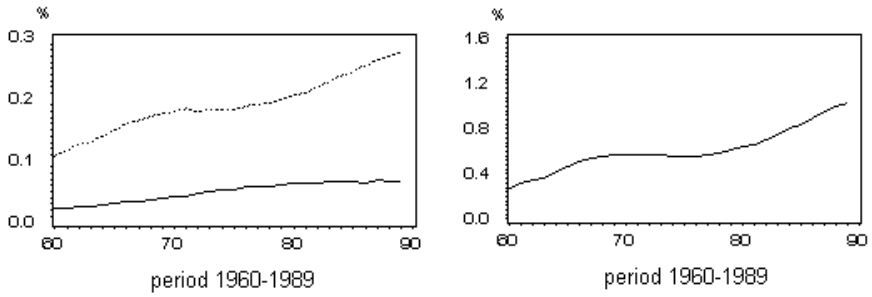


Figure 41.1: Observed IDDM prevalence among females (left: middle, young with solid line; right: old)

The mathematical model chosen was to understand the course of disease prevalence as a stochastic process. This description allows for the integration of population dynamic elements in the mathematical model. The aging population structure of the former GDR is included in the stochastic model, because the disease is dependent on the aging structure. The necessary data is taken from the statistical yearbooks of the GDR (1960 - 1989).
Discrete stochastic processes are easiest to work with. We applied a homogenous, irreducible, ergodic Markov chain with a finite state space.
Because maximum likelihood estimators can be determined for the transition probabilities of such Markov chain, it is possible to use characteristics of these estimators to good purpose in modelling the time course of diabetes mellitus prevalence. Their mathematically proven main properties are asymptotic consistency and asymptotic efficiency. That means roughly spoken, the bias of the calculated transition probabilities tends for large observations to zero, and the information contained in the data is best used.
The database evaluated is extraordinary large. It covers the observation of about 17 million people over 30 years. Consequently, the calculated parameters of the Markov chain are free of bias, and there is no better way to calculate them than applying the maximum likelihood method.
Applying ergodic Markov chains to the modelling of diabetes the stationarity of the stochastic process is the basis of two conclusions. First, there is an equilibrium state of the prevalence in the population. Second, actual observations allow a prognosis of the further development of the diabetes prevalence.
The state space of the Markov chain applied here consists of ten states: healthy (H), IDDM (T1) and NIDDM (T2) in the three age groups and, additionally, the state "gone". Absorbing states must be excluded and a kind of "circu-

lation" introduced. The births belonging to a population development model must therefore be brought back from the state "gone" into the system.
The reader must bear in mind that this is strictly a mathematical aid. This allows us to consider a rather simple model with outstanding characteristics, especially since only the population numbers in each age group are to be considered and the number of individuals in the additional state "gone" will not play a role in the application of the model.
As a starting value for the state "gone", the sum of individuals already present in the other states will be entered. This seems a reasonable start, because then the correct population birth rate can immediately used as a transition probability from the state "gone" to the state "healthy female young".

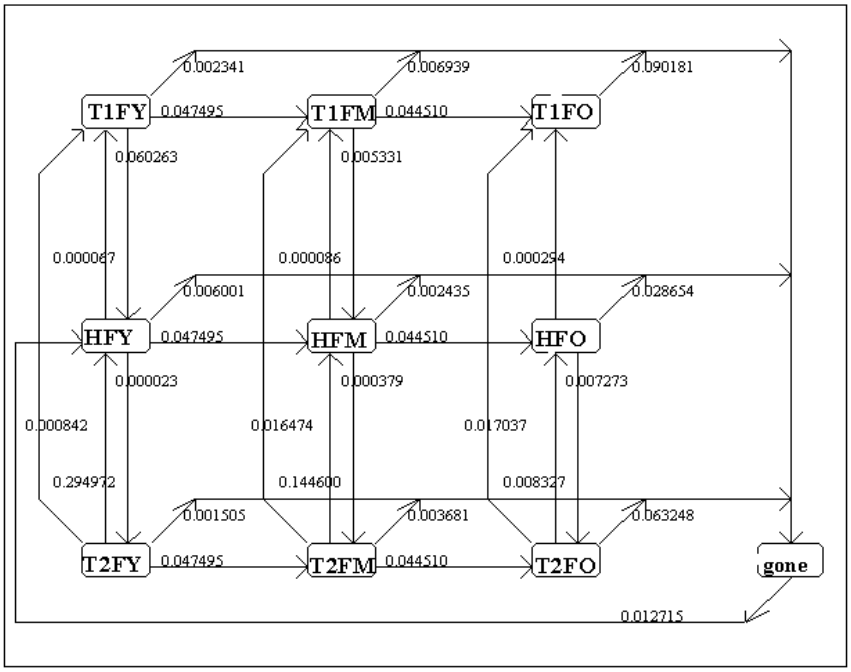


Figure 41.2: Markov chain model for diabetes prevalence including population development, inside: estimated transition probabilities from the data of the GDR diabetes register and the Statistical Yearbooks of the GDR

41.3 Results

The Fig. 41.1 is intended to give an overview of the observed course of IDDM diabetes prevalence in the female subpopulation of the GDR during the years examined. The ordinate axes are divided into percent points. To illustrate

the trend, the individual points are connected by lines. Here must be emphasized that the data from the Karlsburg register for each year only yield one number, and thus the representation should only be limited to 30 single points each. The percent values refer to the proportion of individuals with diabetes mellitus within each age group. Fluctuations within the age structure of the population thus do not influence the values. Next, population development and both treatment forms IDDM and NIDDM will be described for the female population in one Markov chain model (Fig. 41.2). The accompanying transition probabilities form a 10-by-10 field transition matrix. They are estimated by the maximumlikelihood- method and may found in Fig. 41.2. With this model, the prevalence curves ca be simulated over the observed time period. The simulated IDDM prevalences in the female subpopulation are shown for the three age groups in Fig. 41.3.
One sees, observations (cp. Fig. 41.1) and calculations agree well. This impression is confirmed with a goodness-of-fit-measure, the measure of certainty B^2. It is between 0.86997 (group of old females) and 0.99998 (group of young females). Further, the associated asymptotic standard deviation of the calculated transition probabilities reaches at most 50% of the estimated parameter, which is still acceptable. This maximal value concerns, in fact, the smallest observed individual group. It shows only 4753 entries from disease-years, which would be expected in these quite rare cases of young woman treated with oral antidiabetics. Looked at in this way, the model meets the demands on it well.

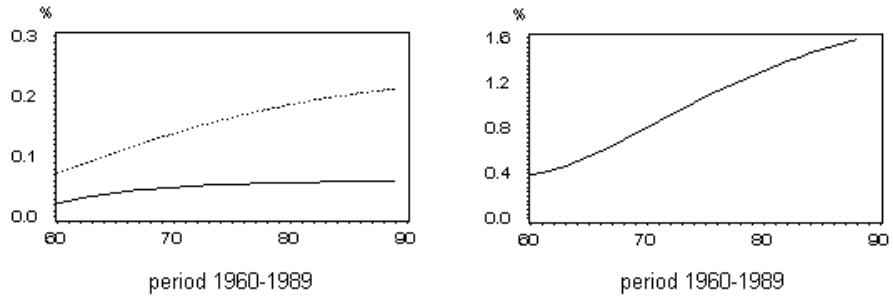


Figure 41.3: IDDM prevalence among females simulated by the Markov chain (left: middle, young with solid line; right: old)

However, we must bear in mind that working with this aspect, the possibilities for refining the model are nearly exhausted: the smaller number of individuals in the single groups or states used to estimate parameters would also let the differences of standard deviations and asymptotic standard deviations increase further.
The observation of the stationary distribution of the Markov chain can be seen as a long-term prognosis for the actual development of the diabetes prevalences.

Independent of the starting distribution, the stationary distribution already indicates that the observed trends can be realized with the model. If the starting distribution is also included, the following trends result:

1. The IDDM prevalences in the two younger female groups attain their stationary values approximately within the 30-year observation period. For the subsequent years, the model predicts only a very slight increase, which in contrast to the previous 30 years does not represent a multiplication.

2. Of special interest is the models prognosis for the further development of IDDM prevalence in the group of older women, which is very strongly influenced by secondary failure. This prevalence curve, derived from the model, is shown in Fig 41.4 . Increasing prevalence values are again prog-

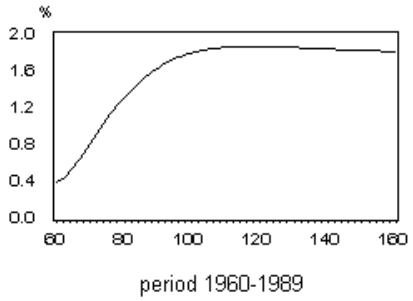


Figure 41.4: IDDM prevalence prognosis among females (old); period 1960 - 2060

   nosed beyond the observed period, with an end to the increase predicted around the year 2010. Here, the prevalence value already lies above the value which follows and represents the stationary distribution. In this case, the model prognoses fluctuations gradually decreasing in towards a constant prevalence value.

3. This models NIDDM prevalences in the young female group demonstrate almost no fluctuations in the 30-year observation period. This trend continues and reaches a stationary value only slightly higher than the starting value.

4. In the middle age group, we again see a prevalence curve of gradually decreasing fluctuations for NIDDM. However, the increase in this case already occurs within the first half of the observation period, then reaches a rather constant, high level, and finally slumps off somewhat during the prognosis time period to attain the stationary distribution value.

5. In the older group of women with NIDDM, a fluctuating-in is also prognosed. A peak is reached approximately at the end of the observation

period; subsequently, the prevalence values decrease somewhat and approach the stationary value.

The model curves found using the male subpopulation data behave as they did in the female subpopulation. The stationary values are approximately reached during the 30-year observation period.

References

1. Gilbertson DT, Liu J, Xue JL, Louis TA, Solid CA, Ebben JP, Collins AJ (2005). Projecting the number of patients with end-stage renal disease in the United States to the year 2015, *J Am Soc Nephrol.*, **16(12)**, 3736–41.

2. Hauner H, von Ferber L, Köster, I (1992). Schätzung der Diabeteshäufigkeit in der Bundesrepublik Deutschland anhand von Krankenkassen. Sekundärdatenanalyse einer repräsentativen Stichprobe AOK-Versicherter der Stadt Dortmund. *Deutsche Medizinische Wochenschrift*, **117**, 645–650.

3. Honeycutt AA, Boyle JP, Broglio KR, Thompson TJ, Hoerger TJ, Geiss LS, Narayan KM. (2003). A dynamic Markov model for forecasting diabetes prevalence in the United States through 2050, *Health Care Manag Sci*, **6(3)**, 155–64.

4. Michaelis D, Jutzi E. (1991). Epidemiologie des Diabetes mellitus in der Bevölkerung der ehemaligen DDR: Alters- und geschlechtsspezifische Inzidenz- und Prävalenztrends im Zeitraum 1960-1987, *Zeitschrift für klinische Medizin*, **46**, 59–64.

5. Müller UA, Ross IS, Klinger H, Geisenheiner S. (1993). Quality of centralized diabetes care: A population-based study in the German Democratic Republic 1989-1990, *Acta Diabetologica*, **30**, 166–172.

6. Salome Ch. (1994). Modellbildung zur Epidemiologie des Diabetes mellitus. Diplomarbeit. Greifswald/Heidelberg.

7. Statistische Jahrbücher der DDR. (1960–89). Berlin

42

Pneumoconiosis Revisited: Classifiers Viewed via ROC Curves and Logic Functions

T. Cacoullos and M. Pattichis

University of Athens
University of New Mexico

Abstract: Region of interest ratings (six-dimensional Bernoulli variables) of 158 chest radiographs of miners are used for their classification into one of two pneumoconiosis q-categories. Six classifiers: Logistic, Bayes, Normal Bayes, K-means, Sum and Weighted Sum, of which 5 were considered in a preceding paper, are compared based on ROC curves, and their performance is evaluated by using logic functions for their representation as binary classifiers. Specifically, Karnaugh maps are constructed, showing lung symmetry and disease growth properties. For each classifier, the area under the ROC curve is estimated and used as a measure of its performance.

43

Bootstrapping Based Inference for a Small Sample Problem from Neonatology

Remus Campean

University of Medicine and Pharmacy, Faculty of Pharmacy, Department of Mathematics and Informatics, Cluj-Napoca, Romania

Abstract: Statistical inferences based on small dimension samples represents a big problem and a made to measure challenge. In the biomedical domain there are numerous situation where costs or ethical reasons enforce that only a few data are collected. Nevertheless, some inferences must be made. In this paper, an alternative model is tested by applying a simulation strategy through a bootstrap re-sampling technique. Linked to this methodology the phenomenon of bootstrap aggregation is revealed. The model is tested on real small samples of data, significant in neonatology. All calculations are implemented through Matlab scripts.

Keywords and phrases: bootstrapping small samples, bagging, correction of bagging, oxidizing stress at newborns

43.1 Introduction

Classical statistics is no longer the only way to infer from data. There are many situations when the conditions for applying a classical tests are not satisfied. The most common inconvenient is the small dimension of the samples and the absence of the information about the population's distribution from which the samples are taken. To handle such situations re-sampling techniques is one of the alternatives. The first part of the paper is dedicated to the presentation of the problem and to a brief introduction into the principles of nonparametric bootstrap re-sampling technique. The second part contains the simulation of an important re-sampling phenomenon: the bootstrap aggregation ("bagging"). In the last part the bootstrapping-simulation model is tested on real small samples of data. These comes from a neonatology problem. The inferential problem is detailed in this part. The interpretation of the results will take into account

the phenomenon of bagging. This is the reason why the simulation of bagging was made before. A correction of bagging through simulation is also discussed here.

Acknowledgements. Data on which the model is tested are obtained from a clinical study concerning the evaluation of the oxidative stress to premature newborns comparing with on time newborns,performed in the neonatology clinical section, the Clinic Obstetric-Gynecology I, from Cluj-Napoca, Romania, under the coordination of professor Antonia Popescu. Superoxide dismutase activity and hemoglobin concentration were determined at the Department of Pharmaceutical Biochemistry and Clinical Laboratory from the Faculty of Pharmacy, Cluj-Napoca by Lecturer Cristina Gagyi. [4]

43.2 The problem and the principle of approach

43.2.1 The statistical inferential problem.

Two samples of small dimension are considered. The first one, X^5=[495.63, 852.07, 468.67, 420.03, 480.2] is sampled from a population of in time newborns, X. The second one, Y^5=[378.71, 337.83, 489.71, 422.29, 520.99] is sampled from a population of premature newborns, Y. The values of X^5 and Y^5 represents doses of superoxide dismutase in both groups [4]. The sample mean of X^5 must be compared with sample mean of Y^5 and to infer that $\overline{X} > \overline{Y}$ with 0.25. Usually, the t test would be enough to prove this, but there is a big problem. The dimensions of the two samples are very small, $n = 5$, and no assumptions on the distributions can be made, especially on Y. In order to infer, alternative methods must be used.

43.2.2 The principle of nonparametric bootstrap re-sampling.

It is not the intention of this paper to present a description of this relatively new statistical technique. However, the principle of *the nonparametric bootstrap re-sampling* is enunciated here. Re-sampling aims to re-construct the distribution of a population starting from one or some selected real samples. The principle of bootstrap re-sampling is *sampling with replacement from the real original sample*. On the basis of these pseudo-samples a new distribution is built. This is called *bootstrap distribution*. From the four variants of bootstrap re-sampling methods (nonparametric, parametric, smoothed and Bayesian) the nonparametric one is used in this statistical study. [1,2]

Observation. To be reliable, this algorithm must be tried for a large number of times, $b = 1..B, B$ of order $10^3, 10^4, ...$ In the sequel, an important phenomenon appears that must be explored, namely, *the bootstrap aggregation*. The phenomenon is studied through simulation in the next section. [3]

43.3 Simulation of the bootstrap aggregation

Bootstrap aggregation is commonly called in short as "bagging". This means the reduction of the variation in the bootstrap distribution comparatively with the real variation. Bagging is observed no matter the distribution of the population is [1,3]. The simulation will underline that the bootstrap distribution tends to a normal shape, with smaller variation, even if the bootstrapped sample doesn't proceed from a normal one. Applying a nonparametric bootstrap algorithm, the Matlab simulation script builds, through B re-sampling iterations, the bootstrap distribution, $BootSampleX$, on the basis of a randomly selected sample, $SampleX$, from the generated distribution X. The statistic applied to each pseudo-sample is the mean.

Simulation: Inputs for bagging from a small sample. Population, $X \sim Exp(\mu)$: $N = 10000$, $\mu = 2.5$; Sample from X, $SampleX$: $n = 10$; Bootstrap distribution, $BootSampleX$: $B = 1000$. Below are shown the numerical results of the simulation and the comparatively distributions: the theoretical one, the sampling and the bootstrap distribution, (Figure 43.1).

Simulated Population:	Simulated Sample:	Bootstrap distribution:
$X \sim Exp(2.5)$	$SampleX$	$BootSampleX$
$\overline{X} = 2.4934 \approx \mu$	$\overline{SampleX} = 2.9913$	$\overline{BootSampleX} = 2.9846$
$S_X = 2.4843$	$S_{SampleX} = 2.5532$	$S_{BootSampleX} = 0.7755$

Observation. There is a small difference between the simulated theoretical mean $\overline{X}$ and the elected parameter $\mu = 2.5$ of the exponential distribution. For the purposes of this simulation this difference is negligible.

In the next application bagging will be corrected in order to make inference more reliable.

43.4 Application: Bootstrapping to compare two small groups of newborns

The inferential problem. The means of the selections X^5 and Y^5 are compared. The tested hypothesis is: the mean of X is 0.25 bigger than the mean of Y. Formally, this means $\frac{\overline{Y}}{\overline{X}} = 0.80$.

43.4.1 The bootstrap algorithm for hypothesis testing

1. A number of B re-samples are constructed from each selections, X^5 and Y^5, with the same dimensions as the original ones. For this application with n=5 the value of B will be taken as $B = 3125 = n^n = 5^5$.

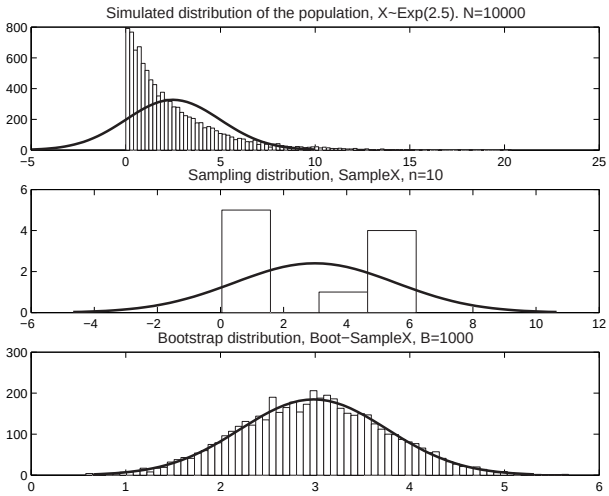


Figure 43.1: **Bagging from a small sample**

2. For each pair of re-samples (X_b^5, X_b^5) the ratio $\frac{\overline{Y_b^5}}{\overline{X_b^5}}$ is calculated. This is the statistic of interest for comparing the groups X and Y.
3. From the b iterations, $b = 1..B$ results the bootstrap distribution of the statistic calculated at point (2), see Fig.43.1.
4. A confidence interval for the bootstrap distribution's mean is calculated.
In the last part of the algorithm will be necessary to take into account the phenomenon of bootstrap aggregation. The importance of the simulation of bagging will be revealed.

Numerical results. By implementing the algorithm above through a Matlab script, the bootstrap distribution of the statistic $\frac{\overline{Y_b^5}}{\overline{X_b^5}}$, b=1..B, is obtained. Its shape is similar with the one obtained in Fig.43.1. For concision, the elements of this distribution will be referred as Boot(Y/X), $\{\frac{\overline{Y_b^5}}{\overline{X_b^5}} | b = 1..B\} =^{not} Boot(Y/X)$. After bootstrapping the statistic of interest, the mean of the distribution is calculated. This will be called *TestBoot*, $\overline{Boot(Y/X)} =^{not} TestBoot = 0.8029$. The confidence interval for TestBoot is calculated on the basis:$Boot(Y/X) \sim N(TestBoot, S_{Boot(Y/X)})$. For unknown theoretical mean μ and the standard deviation calculated from the bootstrap distribution, Boot(Y/X), the confidence interval for μ is considered as: $TestBoot \pm 1.96 \cdot \frac{S_{Boot(Y/X)}}{\sqrt{B}}$. The value 1.96 is the quantile $z_{\alpha/2}$ of the standard normal distribution for a confidence level $\alpha = 0.05$. For the standard deviation of the bootstrap distribution, called *bootstrap standard deviation*, is obtained the value $S_{Boot(Y/X)} = 0.113$. Thus, the confidence interval for *TestBoot* is $TestBoot \in [0.7989, 0.8068]$

43.4.2 Correcting bagging through simulation.

Considering the simulation of bagging experienced in the previous section, the bootstrap aggregation is expected to be present in this distribution too. In order to make bootstrapped inference more reliable a correction of the bootstrap standard deviation would be an appropriate solution. This can be made by studying the variation of $S_{Boot(Y/X)}$ when the bootstrap algorithm is repeated for a large number of times (see Fig.43.2).
Result: $S_{Boot(Y/X)} \in [0.11, 0.1182]$. This is a narrow interval but this reflects the stability of the bootstrap technique. Correction: $S_{Boot(Y/X)}$ can be corrected to a value between 0.11 and $0.2872 = S(\frac{Y^5}{X^5})$. The corrected confidence interval for $TestBoot$ is obtained for the bootstrap standard deviation taken as $2 \cdot max[S_{Boot(Y/X)}] = 2 \cdot 0.1182 = 0.2364$: $[0.7984, 0.8150]$.
The more exact information is known about variation in the real sample, the more the process of correcting bagging can be adjusted.

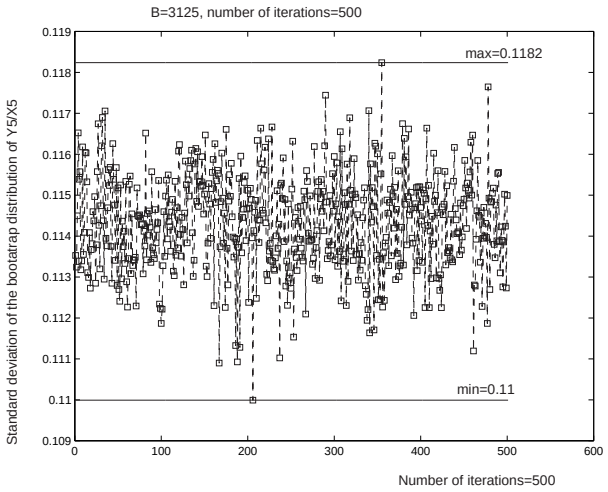


Figure 43.2: **Correction of bagging;** $S_{BootY/X}$ **is calculated** 500 **times for the same** $B = 3125$**.**

43.5 Discussion and Conclusions

Analyzing the results obtained from this approach it can be concluded that the ratio $\frac{\overline{Y}}{\overline{X}}$ in the population is close enough to the expected value 0.80. The argument is that the mean of the re-sampled distribution $Boot(Y/X)$ is very close to this value, $TestBoot = 0.8029$.
The confidence interval, for a good significance level $\alpha = 0.05$, is a narrow one,

which strengthen the idea that the mean of the bootstrap distribution is a good estimator for the theoretical mean of the population.
The phenomenon of bootstrap aggregation must be taken into account. In the majority of problems the study of variation is needed, so the bootstrap aggregation must be corrected.
In several cases bootstrap simulation can be a valid alternative to classical methods.
For this kind of approach, computer implementation is compulsory due to the large number of calculations.
Inferring from small samples is a common problem for many biomedical problems. In these situations classical statistical methods can be replaced with new ones. The bootstrap re-sampling method is one of these alternatives.

References

1. Hesterberg, T. (2003). *Bootstrap Methods and Permutation Tests*, W. H. Freeman and Company, New York.

2. Cheng, R. C. (2000). Analysis of Simulation Output by Re-sampling, *International Journal of Simulation*, Vol 1, **1-2**, 51–58.

3. Schäfer, J. and Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, *Statistical Applications in Genomics and Molecular Biology* (Ed., N. Balakrishnan), Vol 4, Issue 1, article 32.

4. Popescu, A., Matyas, M. and Gagyi, C. (2005). The role of superoxide dismutase in anti-oxidizing defence to newborns - preliminary data, *Proceedings of the 6th National Conference of Perinatal Medicine, Targu Mures, Romania*, pp. 234–237.

44

The Dynamic of Stimulating and Inhibiting Effects in Tissue Culture

N. I. Chalisova, A.A. Chalisova, G. Haase

Pavlov Institute of Physiology, Russian Academy of Sciences
Praxis für Immunotherapie, Germany

44.1 Extended Abstract

Repair processes in tissues are known to be regulated by cytokines in opposite directions: cell proliferation or apoptosis (4, 6, 7). Organotypic tissue culture is successfully used to evaluate rapidly and quantitatively the effect of a substance to be tested (3, 9). The dose dependence of the effect of cytokines usually fit a non-linear dome-shared curve. In other words, the stimulating effects of these agents are observed at specific effective concentrations, whereas at both lower and higher concentrations, the response remains at the control level (1, 2, 5, 8,) The question arises as to what mechanisms underlie stimulating and inhibiting effects observed in tissue culture.

To determine the stimulating and inhibiting dose-dependent effects of various biologically active substances we studied the effect of neurite-stimulating components (destsbilase) secreted by the medicinal leech (effective concentrations 0.05-1 ng/ml) and the effect of the mitogen concanavalin A (effective concentrations 0.01-0.1 mg/ml) on the organotypic culture of the nervous and lymphoid tissues.

It was demonstrated, that at concentrations exceeding the optimal ones, cytokines inhibited cell proliferation when applied alone. In combination with the stimulating agent- nerve growth factor , - the effective concentrations of cytokines had a similar effect. Tissue fragments prepared under sterile conditions were cut to smaller fragments (approximately 1-mm3 pieces) that were placed on Petri dishes containing a collagen substrate. The culture medium contained Eagles basal medium with Earl salt (BME), supplemented with 2 mg/ml glucose, 10 ng/ml insulin, 75 IU penicillin and 25% fetal calf serum.

The control explants were grown in the nutrition medium without additives. The Petri dishes were incubated at 36.5^o C for 48 h and then examined under a phase-contrast microscope. We determined the area index (AI), which was expressed in relative units and calculated as the ratio between the total area of all explant (together with the zone of migrating cells) and the central area of the explant. To detect apoptosis, the explants were stained with 0.1% acridine orange, and their fluorescence intensity was examined under a luminescent microscope. The data were processed using Student's t-test.

In spinal ganglia cultivated with 1 ng/ml destabilase the axon outgrowth was stimulated and AI values increased by 52 ± 5% as compared to the control (n = 23, p< 0.05). However, the higher concentrations of destabilase reduced the growth-stimulating effect. Beginning from a concentration of 2 ng/ml, the explant outgrowth was inhibited. Nerve growth factor added at the concentration 100 ng/ml significantly stimulated the axon outgrowth of sensory neurons: AI values were by 58 ± 11% (n = 22, p< 0.05) higher as compared to the control (n = 25). However, when this neurotrophic factor was used in combination destabilase added at the stimulating concentrations, no stimulation of the explant growth was observed: AI values were by 30 ± 9% (n = 24, p< 0.05) lower as compared to the control (n = 23).

The growth of spleen explants was also significantly stimulated by addition of concanavalin A at concentration 0.1 mg/ml to the culture medium and AI values increased by 158 ± 15% (n = 25, p< 0.05). as compared to the control (n = 21). However, like wit the cultivation of the sensory ganglia, the concanavalin A concentrations higher than the effective once reduced the stimulating effect, which was reversed to induce growth inhibition at concentration higher than 0.4 mg/ml. At the mitogen concentration of 0.5 mg/ml, the AI was by 28 ± 15% (n = 23, p< 0.05) lower as compared to the control (n = 25).

Thus, overstimulation of the tissue cultures "switched off" the proliferative processes, which reflects the cell system adaptability. So the overstimulation provided by the biologically active agents caused the reversion of the response in tissue culture. This problem accounts for the fact that the dose dependencies of the various agents on the tissue cultures are described by dome-shaped curves, rather than monotonous curves with plateauas, because, at large concentrations the stimulating effect of cytokines reverse to inhibition ones. This is a manifestation of one of the generals biological laws: overstimulation caused reversed effect.

References

1. Arai T., Hiromatsu K., Nishimura H., Hamid G. Endogenous interleukin 10 prevents apoptosis in macrophages during Salmonella infection. Biochem. Biophys. Res. Communs. 213 (2): 600-607. 1995.

2. Bing W, Junbao D, Jianguang Q. et al. L-arginine impacts pulmonary vascular structure in rats with an aortocaval shunt . J. Surg Res. 108(1):20-31. 2002.

3. Branton R.L., Clarke D.J. Apoptosis in primary cultures of E14 rat ventral mesencephala: time course of dopaminergic cell death and imlications for neural transplantation. Exp. Neurol., 160 (1): 88-98. 1999.

4. Cid C, Alvarez-Cermeno J.C, Regidor I. et al. Low concentrations of glutamate induce apoptosis in cultured neurons: implications for amyotrophic lateral sclerosis. J.Neurol.Sci.;206(1):91-95.2003.

5. Fratelli M., Gagliardini V., Galli G. et al. Autocrine interleukin-1 beta regulates both proliferation and apoptosis in EL4-6.1 thymoma cells. Blood. 85 (12): 3532-3537. 1995.

6. Fu Y.M, Yu Z.X, Li Y.Q. et al. Specific amino acid dependency regulates invasiveness and viability of androgen-independent prostate cancer cells. Nutr. Cancer.45(1):60-73.2003.

7. Kim K.Y, Moon J.I, Lee E.J. et al. The effect of L-arginine, a nitric oxide synthase substrate, on retinal cell proliferation in the postnatal rat. Dev Neurosci. 24(4):313-21.2002.

8. Llansola M., Bosca L., Felipo V., Hortelano S.Ammonia prevents glutamate-induced but not low K(+)-induced apoptosis in cerebellar neurons in culture. Neuroscience.117(4):899-907.2003.

9. Levi-Montalchini R., Angeletty P. Nerve growth factor. Physiol. Rev.. 48: 534-569. 1982.

45

Protection of Privacy in Randomized Response Techniques

Arijit Chaudhuri, Tasos C. Christofides and Amitava Saha

Applied Statistics Unit, Indian Statistical Institute
Department of Mathematics and Statistics, University of Cyprus
Directorate General of Mines Safety, Dhanbad, Jharkhand, India

Abstract: In estimating the proportion of people bearing a sensitive attribute, following Warner's (1965) pioneering work certain randomized response (RR) techniques are available. These are intended to ensure efficient and unbiased estimation protecting a respondent's privacy when it touches a person's socially stigmatizing feature like induced abortion, testing HIV positive, illegal drug use, etc. Lanke (1976), Leysieffer and Warner (1976), Anderson (1977) and Nayak (1994) among others have discussed how maintenance of efficiency is in conflict with protection of privacy. In their RR-related activities the sample selection is traditionally by simple random sampling with replacement. Following Chaudhuri (2001), here is reported an extension in case of unequal probability sample selection even without replacement.

Observing that multiple responses are feasible in addressing such a dichotomous situation especially with Kuk's (1990) and Christofides' (2003) RR devices, an average of the response-specific jeopardizing measures is proposed.

Keywords and phrases: Efficiency vs privacy, equal and unequal probability sampling, measures of jeopardy, randomized response models

45.1 Introduction

Let $U = (1, \ldots, i, \ldots, N)$ denote a finite population of labelled individuals. Let $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ be a vector of real numbers defined on U as

$$y_i = \begin{cases} 1 & \text{if } i \text{ bears a sensitive attribute } A \\ 0 & \text{if } i \text{ bears the complementary attribute } A^c;\ i \in U. \end{cases}$$

The problem is to unbiasedly and accurately estimate $\theta = Y/N$ or $Y = \sum y_i$ where $\sum$ denotes summation over i in U, on surveying a sample s of units of

U selected with probability $P(s)$ according to a chosen design P. We briefly review two randomized response techniques which can be used for that purpose.

45.1.1 Warner's (1965) randomized response scheme

A box with cards marked A and its complement A^c in proportions $p : (1-p)$, $0 < p < 1$, is offered to a sampled respondent i and the randomized response is

$$I_i = \begin{cases} 1 & \text{if the card type matches the attribute } A \text{ or } A^c \\ 0 & \text{else; } i \in U \end{cases}$$

with the random outcome undivulged. Writing E_R, V_R as operators for expectation and variance with respect to the RR device, we have,

$$E_R(I_i) = py_i + (1-p)(1-y_i) = Prob(I_i = 1)$$
$$V_R(I_i) = E_R(I_i)\,(1 - E_R(I_i)) = p(1-p), \quad i \in U.$$

Then for $r_i = [I_i - (1-p)]\,/(2p-1)$, on ensuring $p \neq \frac{1}{2}$, $E_R(r_i) = y_i$ and $V_i = V_R(I_i) = [p(1-p)]/(2p-1)^2$, $i \in U$.

45.1.2 Kuk's (1990) randomized response technique

A sampled person i reports f_i which is the RR, namely the number of red cards found in k ($k \geq 1$) random draws with replacement from either a box with red and black cards in proportions $p_1 : (1-p_1)$ if the respondent bears A or from a second box if he/she bears A^c in which these proportions are $p_2 : (1-p_2)$ with $0 < p_i < 1$, $i = 1, 2$. Then

$$E_R(f_i) = k[p_1 y_i + p_2(1-y_i)],$$

$$V_R(f_i) = k[p_1(1-p_1)y_i + p_2(1-p_2)(1-y_i)].$$

For $r_i = \left(\frac{f_i}{k} - p_2\right)/(p_1 - p_2)$, ensuring $p_1 \neq p_2$, $E_R(r_i) = y_i$ and

$$V_i = V_R(r_i) = \begin{cases} \frac{p_1(1-p_1)}{k(p_1-p_2)^2} & \text{if } y_i = 1 \\ \\ \frac{p_2(1-p_2)}{k(p_1-p_2)^2} & \text{if } y_i = 0. \end{cases}$$

We should note that the variance V_i depends heavily on the parameters of the randomization device and in essence could be regarded as one of the technical aspects of the device. We may also note that the variance of an estimator for θ involves V_i and increases as V_i itself increases too. It is therefore appropriate to examine the behavior of V_i in relation to the device dependent measure of jeopardy to be introduced n Section 45.3.

45.2 Protection of Privacy

Lanke (1976), Leysieffer and Warner (1976), Anderson (1977) and Nayak (1994) along with many others, in particular with Warner (1965), Kuk (1990), Greenberg et al (1969), and Christofides (2003) confined to SRSWR of the respondents from the population. Under SRSWR, let $P(y = 1) = \theta = Y/N = P(A)$, be the probability that a person chosen from U at random bears the sensitive attribute A. By letting $Prob(\text{Yes}|A) = a$ and $Prob(\text{No}|A^c) = b$ Nayak (1994) noted that

$$Prob(A|\text{Yes}) = \frac{\theta a}{\theta a + (1-\theta)(1-b)}, \quad Prob(A|\text{No}) = \frac{\theta(1-a)}{\theta(1-a) + (1-\theta)b}.$$

Departures of $Prob(A|\text{Yes})$ from θ and $Prob(A^c|\text{No})$ from $1-\theta$ could be treated as measures of revelation of secrecy. Treating R as a response "Yes" or "No", and writing $\text{Prob}(\cdot|\cdot)$ as $P(\cdot|\cdot)$,

$$P(A|R) = \frac{\theta P(R|A)}{\theta P(R|A) + (1-\theta)P(R|A^c)}$$

and

$$P(A^c|R) = \frac{(1-\theta)P(R|A^c)}{(1-\theta)P(R|A^c) + \theta P(R|A)}$$

could be respectively regarded as "revealing probabilities" in announcing R about a person's response concerning A or A^c. If $P(A|R) > \theta$, R is jeopardizing with respect to A and if $P(A^c|R) > (1-\theta)$, then R is jeopardizing with respect to A^c. Combining these two,

$$J(R) = \frac{P(A|R)/\theta}{P(A^c|R)/(1-\theta)}$$

is treated as a "measure of jeopardy" inherent in a response R concerning A or A^c. The higher its value is, the more its deviation from "unity".

The previous discussion applies to the case of selecting the respondents with simple random sampling with replacement. To cover sampling of respondents from U by arbitrary probabilities we proceed in the following way.

Suppose L_i $(0 < L_i < 1)$ to be the probability that y_i is assigned the value "1" for the unit labelled i according to a certain probability mechanism which need not be further specified. Let $L_i(R)$ denote the "conditional probability" that the ith respondent has the stigmatizing characteristic given that his/her randomized response is R. Then

$$J_i(R) = \frac{L_i(R)/L_i}{[1 - L_i(R)]/(1 - L_i)}, \quad i \in U$$

will be defined as the "response-specific" "jeopardy measure" for the RR obtained as R from respondent i. However, since a measure of jeopardy quantifies the risk of revealing his/her status (i.e., whether he/she belongs to the stigmatizing group)which a person undertakes by agreeing to use the randomization device, it should be made known to the participants before they agree to participate in the survey, i.e, before any response is available. It is therefore justified to use a measure which is not response-specific but rather could be regarded as a technical characteristic of the device. We propose as an alternative measure of jeopardy the quantity $\bar{J}_i$, i.e., the arithmetic average of $J_i(R)$ over the alternative forms of R for a given i. The closer $\bar{J}_i$ is to unity, the more the privacy is protected. In addition, $\bar{J}_i$ does not depend on L_i.

45.3 Measures of Jeopardy

For the two RR models, $L_i(R)$, $J_i(R)$ and $\bar{J}_i$ in terms of the relevant parameters are now presented. The quantity $L_i(R)$ is given only for a specific value of R. For example, for Warner's model we give only $L_i(1)$. The calculations for $L_i(0)$ are omitted for reasons of brevity.

45.3.1 Warner's model

Observe that $L_i(1) = pL_i/\left[(1-p) + (2p-1)L_i\right]$. As $p \to \frac{1}{2}$, $L_i(1) \to L_i$ as is desirable for privacy to be protected but $V_i = V_R(r_i) \to \infty$, destroying efficient estimation. In addition, $J_i(1) = p/(1-p)$, $J_i(0) = (1-p)/p$. Thus,

$$\bar{J}_i = \left[J_i(1) + J_i(0)\right]/2 = \left[p/(1-p) + (1-p)/p\right]/2$$

and $J_i(1) = J_i(0) = \bar{J}_i = 1$ if $p = \frac{1}{2}$.

45.3.2 Kuk's model

$$L_i(f_i) = \frac{L_i\left[{p_1}^{f_i}(1-p_1)^{k-f_i}\right]}{{p_2}^{f_i}(1-p_2)^{k-f_i} + L_i\left[{p_1}^{f_i}(1-p_1)^{k-f_i} - {p_2}^{f_i}(1-p_2)^{k-f_i}\right]}.$$

As $p_1 \to p_2$, $L_i(f_i) \to L_i$ but $V_i \to \infty$. In addition,

$$J_i(f_i) = \left[{p_1}^{f_i}(1-p_1)^{k-f_i}\right] / \left[{p_2}^{f_i}(1-p_2)^{k-f_i}\right]$$

with $J_i(f_i) = 1$ if $p_1 = p_2$. Thus $\bar{J}_i = \frac{1}{k+1}\sum_{f_i=0}^{k} J_i(f_i)$ with $\bar{J}_i = 1$ if $p_1 = p_2$.

45.4 Concluding Remarks

Since for the RR models illustrated, the parameters p, p_1, p_2, k are the determining factors for the criteria for protection of privacy, we should specify their values as far as practicable to keep $L_i(\cdot)/L_i$, $J_i(\cdot)$, $\bar{J}_i$ correspondingly close to unity. Trying alternative values of y_i, it is possible to check if the V_i's also may be kept in check. The following tables showing the values of L_i, $L_i(\cdot)$, $J_i(\cdot)$, $\bar{J}_i$, V_i for various design parameters provide a handy guidance.

Table 1 (Warner's Model)
Values of $L_i(1)$, V_i, $J_i(1)$, $J_i(0)$, $\bar{J}_i$ for various values of p and L_i

L_i	0.2	0.4	0.5	0.7				
p	$L_i(1)$				V_i	$J_i(1)$	$J_i(0)$	$\bar{J}_i$
0.44	0.164	0.343	0.440	0.647	17.111	0.785	1.272	1.029
0.52	0.213	0.419	0.520	0.716	156	1.083	0.923	1.003
0.59	0.264	0.489	0.590	0.770	7.466	1.439	0.694	1.066
0.71	0.379	0.620	0.710	0.851	1.167	2.448	0.408	1.428

Table 2 (Kuk's Model)
Values of $L_i(f_i)$, V_i, $J_i(f_i)$ $\bar{J}_i$ for various values of p_1, p_2 and L_i

L_i		0.2	0.4	0.5	0.7	V_i		
p_1	p_2	$L_i(f_i)$				$y_i = 1$	$y_i = 0$	$J_i(f_i)$
$k = 2, f_i = 0$								
0.65	0.90	0.753	0.890	0.924	0.966	1.820	0.720	12.250
0.52	0.58	0.246	0.465	0.566	0.752	34.666	33.833	1.306
0.40	0.22	0.128	0.282	0.371	0.579	3.703	2.648	0.591
$k = 2, f_i = 1$								
0.65	0.90	0.387	0.627	0.716	0.855	1.820	0.720	2.527
0.52	0.58	0.203	0.405	0.506	0.705	34.666	33.833	1.024
0.40	0.22	0.259	0.482	0.583	0.765	3.703	2.648	1.398
$k = 2, f_i = 2$								
0.65	0.90	0.115	0.258	0.342	0.548	1.820	0.720	0.521
0.52	0.58	0.167	0.348	0.445	0.652	34.666	33.833	0.803
0.40	0.22	0.452	0.687	0.767	0.885	3.703	2.648	3.305
$k = 3, f_i = 0$								
0.59	0.73	0.466	0.700	0.777	0.890	4.113	3.352	3.501
0.50	0.44	0.151	0.321	0.415	0.624	23.148	22.814	0.711
0.52	0.69	0.481	0.712	0.787	0.896	2.878	2.467	3.712
$k = 3, f_i = 1$								
0.59	0.73	0.317	0.554	0.650	0.813	4.113	3.352	1.863
0.50	0.44	0.184	0.376	0.475	0.678	23.148	22.814	0.905
0.52	0.69	0.311	0.546	0.643	0.808	2.878	2.467	1.806
$k = 3, f_i = 2$								
0.59	0.73	0.198	0.398	0.497	0.698	4.113	3.352	0.991
0.50	0.44	0.223	0.434	0.535	0.729	23.148	22.814	1.152
0.52	0.69	0.180	0.369	0.467	0.672	2.878	2.467	0.879
$k = 3, f_i = 3$								
0.59	0.73	0.116	0.260	0.345	0.551	4.113	3.352	0.527
0.50	0.44	0.268	0.494	0.594	0.773	23.148	22.814	1.467
0.52	0.69	0.096	0.221	0.299	0.499	2.878	2.467	0.428

Table 3 (Kuk's Model) Values of $\bar{J}_i$ for the values of p_1, p_2 of Table 4 and for $k = 2$ and $k = 3$

k		2	3
p_1	p_2	$\bar{J}_i$	
0.65	0.90	5.099	
0.52	0.58	1.044	
0.40	0.22	1.764	
0.59	0.73		1.720
0.50	0.44		1.058
0.52	0.69		1.706

References

1. Anderson, H. (1977): Efficiency versus protection in a general RR-model. *Scandinavian Journal of Statistics* **4**, 11-19.

2. Chaudhuri, A. (2001): Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *Journal of Statistical Planning and Inference* **94**, 37-42.

3. Christofides, T.C. (2003): A generalized randomized response technique. *Metrika* **57**, 195-200.

4. Greenberg, B.G., Abul-Ela E.L.A., Simmons, W.R. and Horvitz, D.G. (1969): The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association* **64**, 520-539.

5. Kuk, A.Y.C. (1990): Asking sensitive questions indirectly. *Biometrika* **77**, 436-438.

6. Lanke, J. (1976): On the degree of protection in randomized interviews. *International Statistical Review* **44**, 197-203.

7. Leysieffer, R.W. and Warner, S.L. (1976): Respondent jeopardy and optimal designs in RR models. *Journal of the American Statistical Association* **71**, 649-656.

8. Nayak, T.K. (1994): On randomized response surveys for estimating a proportion. *Communications in Statistics, Theory and Methods* **23**, 3303-3321.

9. Warner, S.L. (1965): Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**, 63-69.

46

On Solving Statistical Problems for the Stochastic Processes by the Sufficient Empirical Averaging Method

Evgeniy Chepurin, Alexander Andronov and Asaf Hajiyev

Moscow State University, Vorobyovi Gori, 119992, Moscow, Russia
Riga Technical University, 1 Kalku Str., LV-1658, Riga, Latvia
Institute of Cybernetic, 9 F.Agayev Str., AZ1141, Baku, Azerbaijan

Abstract: A problem of the statistical estimation of the stochastic process functionals is considered. The Sufficient Empirical Averaging method is used. The method requires the existence of the complete sufficient statistics for unknown parameters. Some examples are considered.

Keywords and phrases: Stochastic process, functional, complete sufficient statistic

46.1 Introduction

The problems of the calculation of optimal points estimates for characteristics of random processes functionals, characteristics of scattering of these estimates, and also estimates observed significance levels for a criteria adequacy of model and experimental data occasionally can be successfully solved by using *Sufficient Empirical Averaging* (SEA) method. The method of obtaining statistical results on the basis of the SEA-method consists of the following steps. Let a statistical model $(Y, \mathcal{B}, \mathcal{P})$ generates a sample data $y \in Y$ and admits a complete sufficient statistic $S(y)$. Here Y is a sampling space, $\mathcal{B} = \{A\}$ is sigma-algebra on Y, $\mathcal{P} = \{P\{A;\theta\}, \theta \in \Theta\}$ is a family of probability measures, Θ is parametric space. It is proposed that y is generated by a probability measure with the unknown parameter θ_0, y is a trajectory of $Y(t)$, $0 \le t \le T$, $Y(t)$ is a random process. Let us define conditional distribution $Q(A; s_0) = P\{A|S(y) = s_0; \theta_0\}$. Note that $Q(\cdot; s_0)$ is free of $\theta_0 \in \Theta$. Suppose also that we can simulate a sequence variants of data $y_1^*, \ldots, y_B^*$, where i.i.d. random variables y_i^* are generated by $Q(\cdot; s_0)$. It is well known that each data variant y_i^* is statistically

equivalent to y. Consider at first problems of unbiased point estimation of $g(\theta_0) = E\{G(T, Y(t) \quad \text{for} \quad 0 \le t \le T; \theta_0)\}$, where $G(T, Y(t),\ 0 \le t \le T)$ is interesting for us functional of $Y(t)$. Let $z(y)$ be an easily calculated unbiased estimator for $g(\theta_0)$ that is $E\{z(y); \theta_0\} = g(\theta_0)$. Then SEA-estimate of $g(\theta_0)$ is

$$\hat{g}_B = B^{-1} \sum_{i=1}^{B} z(y_i^*). \tag{46.1.1}$$

Let $\hat{g}^0(S) = E\{z(y)|S\}$. It is the uniformly minimum variance unbiased estimator of $g(\theta)$. If $V\{z(y); \theta_0\} < \infty$ then Eq.(46.1.1) gives the consistent estimate of $\hat{g}^0(S)$ as $B \to \infty$, $E\{\hat{g}_B(S); \theta\} = g(\theta)$ and

$$V_B\{\hat{g}_B(S); \theta_0\} = V\{\hat{g}^0(S); \theta_0\} + \frac{1}{B}(V\{Z(y); \theta_0\} - V\{\hat{g}^0(S); \theta_0\}) \tag{46.1.2}$$

From Eq.(46.1.2) it is easy to choose B for essential proximity $\hat{g}_B(s)$ and $\hat{g}^0(s)$. Often one can get also unbiased estimator for $V\{\hat{g}^0(s); \theta\}$ and other scattering characteristics of $\hat{g}^0(s)$. Notice that many of the unbiased estimation problems can be solved without difficult calculation of probability measures $Q(\cdot; s_0)$ and $E\{G(T, Y(t) \text{ for } 0 \le t \le T; \theta_0)\}$.

Further, every where it is supposed that

$$z(y) = G(T, Y(t)\ for\ 0 \le t \le T).$$

46.2 Base Model

We consider the labeled random process $Y(t)$ that is determined by the sequence $\{\tau_n, \eta_n\}$, $n = 1, 2, ...$, where $\tau_n \in R_+^1$ is a moment of the process events (of the process jumps), $\eta_n = (\eta_{n,1}, \eta_{n,2}, ..., \eta_{n,m})^T$ is the corresponding label, $\eta_n \in R^m$. Note that a part of η_n can be integers. It is supposed that the sequences $\{\tau_n\}$ and $\{\eta_n\}$ are independent. Furthermore let $K(t) = max\{n : \tau_n \le t\}$ be a number of the process events on the interval $[0, t]$. It is known the sample trajectory of the process $K(t)$ is statistically equivalent to an evolution of the sequence $\{\tau_n\}$ - the jump moments of the process $K(t)$.

Many problems of the queueing theory, reliability, insurance, inventory etc. can be presented as searching problem of the expectation for a functional $G(T, Y(t), 0 < t \le T)$ where T is a fixed time moment. Let θ_0 be a generating parameter of the process $Y(t)$. If its value is unknown then there arises a searching problem of the optimal unbiased estimate for $E\{G(T, Y(t), 0 < t \le T);\ \theta_0\}$. Note that the corresponding unbiased estimate exists for special observation plans about the process $Y(t)$ only. So it exists for the following plans, for example:

- Plan of A-type: the process $Y(t)$ is observed in the interval $\{0, T\}$;

- Plan of B-type: a time moment of observation ending coincides with $\tau_{n(0)}$ where $n(0)$ is such that

$$P\{K(T) \leq n(0);\ \theta_0\} = 1.$$

Unfortunately for the substantial practical problems usually it is impossible to find an analytical expression for the optimal unbiased estimate. On the other hand it is often possible to find the unbiased estimate that is plenty of close to the optimal one. These estimates can be gotten by using the *Sufficient Empirical Averaging* method that has been proposed by Chepurin (1994, 1995, 1999).

46.3 On a Class of Processes with the Complete Sufficient Statistics for the Plans of A-Type

In the current section it is supposed that $\theta_0 = (\theta_{0,1}, \theta_{0,2})$ where $\theta_{0,1}$ determines the distribution of the sequence $\{\tau_n\}$, $\theta_{0,2}$ determines the distribution of the label sequence $\{\eta_n\}$. We suppose that the statistical model generating the process $Y(t)$ admits a complete sufficient statistic $S_1 = K(T)$ for the sequence $\{\tau_1, \tau_2, ..., \tau_{K(T)}\}$. It means that joint probability density of the random sequence $\{\tau_1, \tau_2, ..., \tau_{K(T)}; K(T)\}$ can be represented in the following way:

$$Lc'(\tau_1, \tau_2, ..., \tau_{K(T)}; K(T)) = V(t_1, t_2, ..., t_k) exp\{-\theta_{0,1} k + a_1(\theta_{0,1}) + b_1(k)\},$$

where $V(t_1, t_2, ..., t_k)$ is an arbitrary joint probability density of the vector $\{\tau_1, \tau_2, ..., \tau_k\}$ on the set $0 < t_1 < t_2 < ... < t_k \leq T$.

Here and below $a_i(.)$ and $b_i(.)$ are components of density representation for the one-index exponential family.

Furthermore let S_2 be a restrictedly complete sufficient statistic for the family of the conditional random sequence of the labels $\{\eta_1, \eta_2, ..., \eta_{K(T)} | K(T) = k\}$. It is simply to show that $S = (S_1, S_2)$ is the complete sufficient statistic. As for a structure of $Y^*(t)$ (a date variant for the labeled random process), it is described in the following way: Y*(t) is determined uniquely by the sequence

$$(t_1^*, \eta_1^*), (t_2^*, \eta_2^*), ..., (t_k^*, \eta_k^*),$$

where $(t_1^*, t_2^*, ..., t_k^*)$ are generated by the probability density $V(.)$,

> $(\eta_1^*, \eta_2^*, ..., \eta_k^*)$ are a date variant for the sequence of the labels $(\eta_1, \eta_2, ..., \eta_k)$ provided fixed values of the complete sufficient statistic S_2.

Let us consider an important particular example of the point process, for which $K(t)$ is the complete sufficient statistic.

Example 1. Mixed Poisson process.

Let $K(t), 0 \le t \le T$, be the standard Poisson process with the parameter $\lambda > 0$, $0 \le t \le T$, with that λ is a realization of the random variable Λ with the probability density from the one-index exponential family:

$$Lc'(\Lambda) = exp\{-\lambda/\sigma_0 + a_2(\lambda) + b_2(\sigma_0)\},$$

so $\theta_{0,1} = 1/\sigma_0$.

Let us show that $K(T)$ is the complete sufficient statistic and the conditional probability density $Lc'(\tau_1, \tau_2, ..., \tau_{K(T)}|K(T) = k)$ coincides with the probability density of the order statistic set for a sample from k independent but distributed on $[0, T]$ random variables. Actually

$$Lc'(\tau_1, \tau_2, ..., \tau_{K(T)}|K(T) = k) = \int_0^\infty \left(\prod_{i=1}^k \lambda e^{-\lambda t_i}\right) exp\left\{-\lambda\left(T - \sum_{i=1}^k t_i\right)\right\}$$

$$exp\left\{-\frac{\lambda}{\sigma_0} + a_2(\lambda) + b_2(\sigma_0)\right\} d\lambda = \frac{k!}{T^k} \int_0^\infty \frac{1}{k!}(\lambda T)^k e^{-\lambda T} exp\left\{-\frac{\lambda}{\sigma_0} + a_2(\lambda) + b_2(\sigma_0)\right\} d\lambda.$$

If we take in mind that

$$Lc'(POIS(\Lambda T)) = \int_0^\infty \frac{1}{k!}(\lambda T)^k e^{-\lambda T} exp\left\{-\frac{\lambda}{\sigma_0} + a_2(\lambda) + b_2(\sigma_0)\right\} d\lambda$$

is the unconditional probability density of the random variable $K(T)$ then above formulated statement about the structure of $Lc'(\tau_1, \tau_2, ..., \tau_{K(T)}|K(T) = k)$ becomes obvious.

Note if we set

$$a_2(\lambda) = ln\frac{\lambda^{a_0 - 1}}{\Gamma(a_0)},\ b_2(\sigma_0) = -ln\sigma_0^{a_0},$$

in other words to suppose that Λ has gamma distribution with known form parameter a_0 and unknown scale parameter σ_0, then for the unconditional probability we have the negative binomial distribution:

$$P_S\{K(T) = k;\ \theta_{0,1}\} = \binom{a_0 + k - 1}{k}\left(\frac{1}{\sigma_0 T + 1}\right)^k \left(\frac{\sigma_0 T}{\sigma_0 T + 2}\right)^{a_0}.$$

Let us show the completeness of the unconditional distribution of $K(T)$. Actually let we have $E\{d_{K(T)};\ \theta_{0,1} \equiv 0\}$ for some sequence $\{d_0, d_1, ...\}$ and for all σ_0. Then

$$\sum_{k=0}^\infty d_k \int_0^\infty \frac{(\lambda T)^k}{k!} e^{-\lambda T} exp\left\{-\frac{\lambda}{\sigma_0} + a_2(\lambda) + b_2(\sigma_0)\right\} d\lambda =$$

$$= \int_0^\infty \left(\sum_{k=0}^{\infty} d_k \frac{(\lambda T)^k}{k!} e^{-\lambda T} \right) exp \left\{ -\frac{\lambda}{\sigma_0} + a_2(\lambda) + b_2(\sigma_0) \right\} d\lambda.$$

Now from the completeness of the distribution of the random variable Λ follows that

$$\sum_{k=0}^{\infty} d_k \frac{1}{k!} (\lambda T)^k e^{\lambda T} = 0 \; almost \; probably \; for \; all \; \lambda.$$

In one's turn, from the completeness of the Poisson distribution follows that $d_k = 0$ for $k = 0, 1, ...$, so $K(T)$ is the complete sufficient statistic. Note that an example of the B-type plan has been considered by Andronov *et al.* (2005).

46.4 On Procedures of Data Variant Generation

The problems of a data variants simulation are crucial for possibility of the supposed method realization. To simulate data variant it is necessary to know the conditional distribution of the data variant and to generate corresponding random variables. Usually it is very difficult to find explicit form for the conditional distribution, since it is a distribution on hypersurface in space of high dimension. On the other hand, to generate corresponding random variables is complicated problem too. Here two ways are possible. Firstly, often we can generate the random variables of interest directly, without knowledge of the corresponding distribution. Such examples were given by Chepurin (1995, 1999), Engen and Lillegard (1997).

Secondly, it is possible to apply the *Gibbs sampling*. This approach uses a decomposition of the multivariate probability density into a marginal and then a sequence of conditionals. We begin with the univariate marginal distribution (provided fixed value of the corresponding complete sufficient statistic) and generate the first random variable χ_n^*. Then we recount the value of the statistic and use one for the generation of the next random variable χ_{n-1}^* etc.

We illustrate this approach for a sample $\chi_1, \chi_2, ..., \chi_n$ from the normal population $N(\mu, \sigma)$. In this case the complete sufficient statistic is $S = (\mu_n^*, \sigma_n^{2*})$,

$$\mu_n^* = \frac{1}{n} \sum_{i=1}^{n} \chi_i, \sigma_n^{2*} = \frac{1}{n-1} \sum_{i=1}^{n} (\chi_i - \mu_n^*)^2.$$

The conditional random variable χ_n^* by the condition $S = (\mu_n^*, \sigma_n^{2*})$ has the following probability density:

$$Lc' \left(\chi_n^* | \mu_n^*, \sigma_n^{2*} \right) = \frac{\sqrt{n} \Gamma \left(\frac{n-1}{2} \right)}{(n-1)\sqrt{\pi \sigma_n^{2*}} \Gamma \left(\frac{n-2}{2} \right)} \left(1 - \frac{n}{(n-1)^2 \sigma_n^{2*}} (x - \mu_n^*)^2 \right)^{\frac{n}{2}-2},$$

$$\mu_n^* - \frac{n-1}{\sqrt{n}} \sqrt{\sigma_n^{2*}} \leq x \leq \mu_n^* + \frac{n-1}{\sqrt{n}} \sqrt{\sigma_n^{2*}}.$$

Now we generate χ_n^* using, for example, *Acceptance/Rejection or Inverse Cumulative Distribution Function* methods. Furthermore we recount the value of the statistic S by the formulas

$$\mu_{n-1}^* = \frac{1}{n-1}(n\mu_n^* - \chi_n^*),\ \sigma_{n-1}^* = \frac{n-1}{n-2}\left(\sigma_n^* - \frac{1}{n}(\chi_n^* - \mu_{n-1}^*)^2\right).$$

The consequent iterations give the sequence $\chi_n^*, \chi_{n-1}^*, ..., \chi_4^*, \chi_3^*$. Two last values are calculated by formulas

$$\chi_2^* = \mu_2^* - \sqrt{\frac{1}{2}\sigma_2^{2*}},\ \chi_2^* = \mu_1^* - \sqrt{\frac{1}{2}\sigma_2^{2*}}.$$

Finally it is possible to conclude that the supposed approach allows efficiency to apply the various models of queueing theory, reliability, inventory, insurance etc. for practical problem solving.

References

1. Andronov, A., Zhukovakaya, C. and Chepurin, E. (2005). On Application of the Sufficient Empirical Averaging Method to Systems Simulation. In *Proceedings of the 12th International Conference on Analytical and Stochastic Modelling Technique and Applications*, pp.144–150, Riga, Latvia.

2. Chepurin, E.V. (1994). The Statistical Methods in Theory of Reliability. *Obozrenije Prikladnoj i Promishlennoj Matematiki, Ser. Verojatnost i Statistika*, **Vol.1, N 2**. 279–330. (In Russian.)

3. Chepurin, E.V. (1995). The Statistical Analysis of the Gauss Data Based on the Sufficient Empirical Averaging Method. *Proceeding of the Russian University of People's Friendship. Series Applied Mathematics and Informatics*, **N 1**. 112–125. (In Russian.)

4. Chepurin, E.V. (1999). On Analitic-Computer Methods of Statistical Inferences of Small Size Data Samples. In: *Proceedings of the International Conference Probabilistic Analysis of Rare Events.* (Eds., V.V. Kalashnikov and A.M. Andronov), pp.180–194, Riga Aviation University, Riga.

5. Engen, S. and Lillegrad, M. (1997). *Stochastic Simulations Conditioned of Sufficient Statistics*, Biometrica, **Vol. 84, N 1**, pp. 235–240.

47

Equol Improves the Capacity of Tamoxifen to Prevent Mammary Tumors by Preventing Oxidative DNA Damage

Andreas I. Constantinou, Bethany E. P. White and Katerina Nicolaou

Department of Biological Sciences, College of Pure and Applied Sciences University of Cyprus, Nicosia, Cyprus

47.1 Introduction

In 1998, the National Surgical Adjuvant Breast and Bowel Project (NSABP) demonstrated that TAM treatment reduced the incidence of both invasive and noninvasive breast cancer in women at high risk for the disease(1). Following the results of this study, women at high risk for developing breast cancer are prescribed TAM to prevent breast cancer. TAM is also prescribed to those women with ER positive breast cancer who have undergone chemotherapy to prevent secondary tumors. In recent years, health conscious women from Europe and USA began consuming soy protein or taking soy isoflavones as supplements for their apparent benefits against breast cancer and cardiovascular disease. Consequently, a group of women that are prescribed TAM may also consume soy products or take a mixture of isoflavones, composed mainly of genistein and daidzein, as dietary supplements. The majority of previous studies have focused on genistein, due to its relatively strong binding (in comparison to daidzein) to ER and its estrogenic/antiestrogenic activities which are stronger than those of other isoflavones (2-4). At relatively high concentrations and in vitro, genistein is known to inhibit enzymatic activities that are crucial for tumor cell proliferation (5), such as the receptor tyrosine kinases (6) and topoisomerase II (7-9). Genistein has also been found to be effective in preventing DMBA-induced mammary tumorigenesis in female rats, but only when administered to neonatal or prepubescent rats less than 35 days old (10,11). However, another study found genistein to have no protective effect on DMBA-induced mammary

tumors in mice and even suggested a potentially adverse effect on tumor development when high levels of genistein are consumed (12). Two studies reported no significant differences in the effects of isoflavone-containing or -depleted soy protein isolate in DMBA-induced mammary carcinogenesis (13,14).

47.2 Results

It is unknown at present how the risk of breast cancer is affected by combining TAM with soy phytoestrogens. To address this question, female Sprague-Dawley rats were placed on diets supplemented with TAM, genistein, daidzein, or a combination of each isoflavone with tam; a week later the rats were given the carcinogen 7, 12 dimethylbenz[a]- anthracene (DMBA). The most effective diet was the TAM/daidzein combination: it reduced tumor multiplicity by 76%, tumor incidence by 35%, tumor burden by over 95%, and it increased tumor latency by 62% compared to the control basal diet. The TAM/daidzein combination diet was in all aspects more effective while theTAM/genistein combination was less effective than the TAM diet. The expression of ER in the mammary glands of rats fed daidzein tripled in comparison to those fed the basal diet according to Western data. There was a significant decrease in the expression of ER in the tam-fed rats. The TAM/daidzein diet significantly decreased 8-oxo-dG levels (an indicator of oxidative DNA damage) in the mammary glands. This study conclusively shows for the first time the combination of daidzein with TAM produces increased protection against mammary carcinogenesis, while the combination of genistein with TAM produces an opposing effect when compared to TAM alone. The daidzein metabolite equol may account for the benefit produced by the daidzein diets. The data also suggest that up-regulation of ER may lead to increased protection against carcinogenesis. The effects of these diets in endometrial tumor parameters have been determined (and will be reported here) to determine how equol affects the anticipated adverse effects of tamoxifen in the uterus.

References

1. Fisher, B. et al. TAM for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 90, 1371-88 (1998).

2. Kuiper, G.G. et al. Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor beta. Endocrinology 139, 4252-63 (1998).

3. Zava, D.T. & Duwe, G. Estrogenic and antiproliferative properties of genistein and other flavonoids in human breast cancer cells in vitro. Nutr Cancer 27, 31-40 (1997).

4. Branham, W.S. et al. Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor. J Nutr 132, 658-64. (2002).

5. Kim, H., Peterson, T.G. & Barnes, S. Mechanisms of action of the soy isoflavone genistein: emerging role for its effects via transforming growth factor beta signaling pathways. Am J Clin Nutr 68, 1418S-1425S (1998).

6. Akiyama, T. & Ogawara, H. Use and specificity of genistein as inhibitor of protein-tyrosine kinases. Methods Enzymol 201, 362-70 (1991).

7. Kiguchi, K., Constantinou, A.I. & Huberman, E. Genistein-induced cell differentiation and protein-linked DNA strand breakage in human melanoma cells. Cancer Commun 2, 271-7 (1990).

8. Constantinou, A., Kiguchi, K. & Huberman, E. Induction of differentiation and DNA strand breakage in human HL-60 and K-562 leukemia cells by genistein. Cancer Res 50, 2618-24 (1990).

9. Constantinou, A.I., Mehta, R.G. & Vaughan, A. Inhibition of N-methyl-N-nitrosourea-induced mammary tumors in rats by the soybean isoflavones. Anticancer Res 16, 3293-8 (1996).

10. Barnes, S. The chemopreventive properties of soy isoflavonoids in animal models of breast cancer. Breast Cancer Res Treat 46, 169-79 (1997).

11. Lamartiniere, C.A., Moore, J., Holland, M. & Barnes, S. Neonatal genistein chemoprevents mammary cancer. Proc Soc Exp Biol Med 208, 120-3 (1995).

12. Day, J.K. et al. Dietary genistein increased DMBA-induced mammary adenocarcinoma in wild-type, but not ER alpha KO, mice. Nutr Cancer 39, 226-32. (2001).

13. Constantinou, A.I. et al. Chemopreventive effects of soy protein and purified soy isoflavones on DMBA-induced mammary tumors in female Sprague-Dawley rats. Nutr Cancer 41, 75-81. (2001).

14. Cohen, L.A., Zhao, Z., Pittman, B. & Scimeca, J.A. Effect of intact and isoflavone-depleted soy protein on NMU-induced rat mammary tumorigenesis. Carcinogenesis 21, 929-35. (2000).

48

Fuzzy Based State Reduction Technique for Multi-State System Reliability Assessment

Yi Ding[1,2], **Anatoly Lisniaski**[1,3] **and Ilia Frenkel**[1]

[1]*International Reliability and Risk Management Center (IRRMC) Sami Shamoon College of Engineering, Israel*
[2]*Nanyang Technological University, Singapore*
[3]*Israel Electric Corporation Ltd., Israel*

Abstract: In the multi-state system (MSS) each system element may have many different states. Therefore the computational burden becomes the crucial factor by using multi-state models when there is a "dimension damnation" problem caused by the element state increase. To avoid this problem, the states of a system element are usually reduced into the specified binary- state or three-state model. However the main disadvantage of this approach is that the specified binary- state or three-state model cannot fully represent the performance or/and statistical behavior of all the actual states. In this paper a fuzzy based reduction technique has been developed to cluster system states and simplify computational complexity for a MSS. It can be seen from the illustrative example that the proposed technique is accurate and achieves much better performance than the conventional approximation method.

Keywords and phrases: Fuzzy, reliability, multi-state system, state reduction

48.1 Introduction

There are many technical systems in the world, which are designed to perform their intended tasks in a given environment. One type of these technical systems is Multi-state Systems (MSS). MSSs have a finite number of performance rates (intensity of the task accomplishment). They are able to perform their task with various performance rates. Failures of some system elements only lead to the degradation of the system performance. The basic concepts of MSS reliability were primarily introduced by Murchland (1975), El-Neveihi et al. (1978), and Barlow and Wu (1978). The comprehensive up-to-date presentation of the MSS reliability theory and its applications were discussed by Lisnianski and Levitin (2003).

In the multi-state system (MSS) each system element may have many different states. Therefore the computational burden becomes the crucial factor by using multi-state models when there is a "dimension damnation" problem caused by the element state increase. Some achievements like Billinton and Wee (1985) can drastically reduce the number of states and the computation burden. By using these approaches, the original multi-state model of a system element can be simplified as a revised binary-state model or a revised multi-state model. However by using the revised binary-state model, the performance rate (level) can only function in either a perfect functioning rate (level) or a complete failure, or by using a revised multi-state model, the performance rates (level) of the specified derated states are usually simplified defined as the average value in a specified interval. This kind of simplification may have an inaccurate result to reliability evaluations in some cases. How to optimally determine the specified states that represent the characteristics of original states has not been considered in the previous research.

In this paper a fuzzy based reduction technique has been developed to simplify the computational burden of MSS reliability assessment. In this technique fuzzy-c-means (FCM) algorithm has been used to partition the original states of the system element or the subsystem into different specified clusters according to their characteristics. The method for determining the probabilities and associated rates (level) of the specified clustering states has been proposed. In the illustrative example the computational results are compared with results obtained by the exact method.

48.2 Fuzzy-c-means Algorithm

The fuzzy-c-means (FCM) is a data clustering technique and provides a method to group data points into a specific number of different clusters introduced by Bezdek et al. (1984). In this paper, the FCM technique is used to partition the original states of the system element or the subsystem into different specified clusters according to their characteristics. The specified fuzzy states are used to represent the characteristics of the corresponding clusters.

The objective of FCM algorithm is to minimize the specified function F:

$$\min F = \sum_{i=1}^{N} \sum_{s=1}^{S} (U_{si})^m \left\| g_i - c_s \right\|^2, 1 \leq m < \infty \tag{1.1}$$

where N is the number of original states, S is the number of specified clusters (specified fuzzy states), g_i represents the performance rate (level) of the i^{th} original state, c_s signifies the performance rate (level) of the center of the s^{th} cluster, m is any real number greater than 1, U_{si} is the membership grade, which represents the weighting factor between g_i and c_s, and $|| \star ||$ is any norm expressing the similarity between any measured data and the center.

The objective function F represents the distance from any given g_i to a cluster center c_s, which is weighted by the membership value of g_i. While F is minimized, the N original states can be partitioned into S clusters (fuzzy states). An iteration algorithm was developed by Bezdek (1984) to minimize the F:

1. For each cluster center c_s, guess an initial value; for each g_i and each c_s, initialize the membership grade U_{si}

2. For each iteration h, the c_s and U_{si} can be calculated by the following equations, respectively:

$$c_s^{(h)} = \frac{\sum_{i=1}^{N} (U_{si}^{(h-1)})^m \cdot g_i}{\sum_{i=1}^{N} (U_{si}^{(h-1)})^m} \tag{1.2}$$

$$U_{si}^{(h)} = \frac{1}{\sum_{j=1}^{S} \left[\frac{||g_i - c_s^{(h)}||}{||g_i - c_j^{(h)}||} \right]^{2/(m-1)}} \tag{1.3}$$

3. $\left\| U_{si}^{(h)} - U_{si}^{(h-1)} \right\| \leq \varepsilon$, stop the iteration; else go to step 2; where ε is the specified tolerance level of convergence.

48.3 Proposed Method for MSS Reliability Assessment

Let the polynomial $\sum_{i=1}^{N} p_i \cdot z^{g_i}$ represent the performance distribution of the system element or the system. p_i represents the probability of the i$^{\text{th}}$ original state. Such polynomial will be also called as individual universal generating function (UGF) representation by Lisnianski and Levitin (2003).

The FCM algorithm is used by the operator $\tilde{\phi}_{FS}$ to obtain the specified fuzzy states:

$$\tilde{\phi}_{FS}(\sum_{i=1}^{N} p_i \cdot z^{g_i}) = \sum_{s=1}^{S} p_s \cdot z^{\tilde{g}_s} = \sum_{s=1}^{S} p_s \cdot z^{\{g_i, U_{si} | g_i \in G_i\}} \tag{1.4}$$

where $\tilde{g}_s$ is the performance rate (level) of fuzzy state s, which is represented by a fuzzy value, p_s is the probability of fuzzy state s and G_i are collection of objects denoted generically by g_i. Equation 1.4 is the representation of fuzzy universal generating function (FUGF). p_s can be calculated by the following

equation:

$$p_s = \sum_{i=1}^{N} \left(\frac{U_{si}}{\sum_{s=1}^{S} U_{si}} \cdot p_i \right) \tag{1.5}$$

From the equation 1.5, it can be seen that the calculation of p_s is based on apportioning the probabilities of the original states into the specified fuzzy states. The closer an original state to a cluster center which represents that U_{si} is a relatively high value, the more contribution it gives to the probability of that fuzzy state.

However it can be seen from (1.5) that the performance rate (level) $\tilde{g}_s$ of fuzzy state s is a discrete fuzzy number with N elements, which will result in a heavy computational burden. To reduce the computational complexity, the $\tilde{g}_s$ can be approximated by some kind of continuous membership function. In this paper, it is supposed that $\tilde{g}_s$ can be approximated by a moving up triangular as shown in Fig. 1.

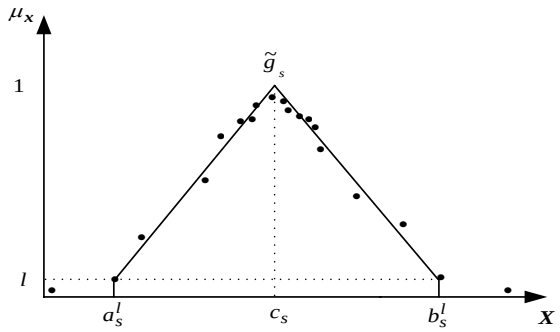


Fig. 1. The approximation of a discrete fuzzy number by a moving up triangular

It is supposed that the interval confidence of $\tilde{g}_s$ at l-cut level $\tilde{g}_s^l$ is determined as $\left[a_s^l, b_s^l\right]$ shown in Fig.1. The membership function of $\tilde{g}_s$ approximation is defined as:

$$\mu_{g_s}^{app}(x) = \begin{cases} 0, & x < a_s^l, \\ \frac{x - a_s^l}{c_s - a_s^l} + l, & a_s^l \le x \le c_s, \\ \frac{b_s^l - x}{b_s^l - c_s} + l, & c_s \le x \le b_s^l, \\ 0, & x > b_s^l \end{cases} \tag{1.6}$$

Especially notice that when l is defined as 1, the $\tilde{g}_s$ approximation is reduced into a crisp value - cluster center c_s, which can be seen as the pseudo performance rate (level) of the specified state s.

The operator $\tilde{\phi}_{app}$ is defined to approximate $\tilde{g}_s$ to a moving up triangular:

$$\tilde{\phi}_{app}(g_i, U_{si}|g_i \in G_i) = (a_s^l, c_s, b_s^l)_l \tag{1.7}$$

When $l = 1$, $\tilde{\phi}_{app}(g_i, U_{si}|g_i \in G_i) = c_s$. The $(a_s^l, c_s, b_s^l)_l$ can be seen as that the triangular represented by the conventional triplet (a_s^l, c_s, b_s^l) moves up a vertical level l.

After obtaining the FUGF of the system element or the system, the mathematical calculations for the FUGF discussed by Ding and Lisnianski (2006) can be used to calculate the system reliability indices.

48.4 Illustrative Example

In the example, the availability of a generation company (Genco) is evaluated by the fuzzy based reduction technique. It is supposed that the Genco has four generating units: two same coal units, one gas unit and one oil unit. The reliability data of generating units come from Israel power system. The coal unit, the gas unit and the oil unit have 10 states, 10 states and 11 states, respectively. Therefore the Genco totally has $10 * 10 * 10 * 11 = 11000$ states.

Obviously the Genco is a parallel system. In the first step, the two coal units are combined into a subsystem 1 and the gas unit and oil unit are combined into a subsystem 2. For each subsystem, the derated states are clustered into 8 fuzzy states by using fuzzy based reduction technique. The characteristics of the subsystem are represented by a state of total failure, a state of full capacity and 8 fuzzy derated states. The fuzzy value of the performance level for the fuzzy derated state is calculated by FCM and approximated by a triangular with a moving up vertical level l. The probability for the fuzzy derated state is evaluated by (1.5). There are only $10*10 = 100$ items in the system FUGF. The system availabilities with different demand levels evaluated by exact method, in which there is no state reduction and all the 11000 system states are evaluated and fuzzy based reduction technique with l level 0.95 and 0.6 respectively are shown in Fig.2.

References

1. J. Murchland (1975). Fundamental concepts and relations for reliability analysis of multi-state systems and fault tree analysis, Theoretical and Applied Aspects of System Reliability, SIAM, pp. 581-618.

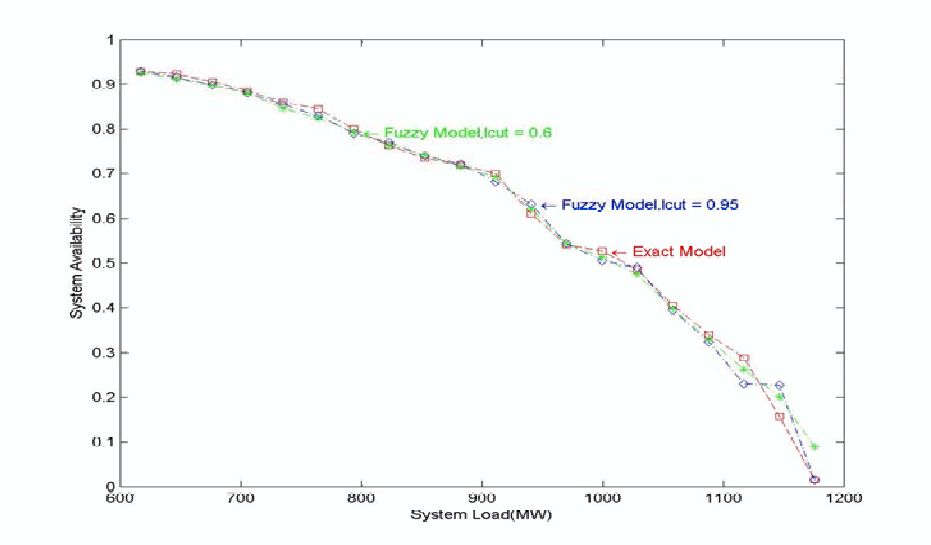


Fig. 2: System availabilities evaluated by exact method and fuzzy method

2. E. El-Neveihi, F. Prochan and J. Setharaman (1978). Multi-state coherent systems, J. Applied Probability, vol. 15, pp. 675-688.

3. R. Barlow and A. Wu (1978). Coherent systems with multi-state elements, Math. Operation Research, vol. 3, pp. 275-281.

4. A. Lisnianski and G. Levitin (2003). Multi-state system reliability Assessment, Optimization, Applications, World Scientific.

5. R. Billinton and C. Wee (1985). Derated State Modeling of Generating Units, Report prepared for Saskatchewan Power Cooperation.

6. J.C. Bezdek, R. Ehrlich and W. Full (1984). FCM: The fuzzy-c-means clustering algorithm, Cmpei. Gcosci., vol. 10, pp. 191-203.

7. Y. Ding and A. Lisnianski (2006). Fuzzy Universal Generating Functions for Multi-state System Reliability Assessment, Submitted to Reliability Engineering and System Safety for Publication.

49

Bayesian Analysis of Correlated 2×2 Contingency Tables

Anastasia G. Eleftheraki[1], Maria Kateri[1] and Ioannis Ntzoufras[2]

[1]*Department of Statistics and Insurance Science, University of Piraeus, Piraeus, GREECE*
[2]*Department of Statistics, Athens University of Economics and Business, Athens, GREECE*

Abstract: The analysis of correlated binary responses for two independent groups is considered here. We focus on the case where the only available information are the marginal 2×2 crosstabulations between a group variable and the response for two separate occasions (usually time sequences) and not the individual responses. Assuming independent binomial distributions in the k-th table, our objective is to estimate the success probabilities for each group at each occasion as well as the corresponding odds ratios θ_k, comparing the responses of the two groups at each occasion k. In order to deal with the missing information of each subject's response and to estimate the corresponding transition probabilities, a Bayesian procedure is adopted.

Keywords and phrases: Repeated binary response, marginal tables, latent individual information, MCMC

49.1 Introduction

Let us consider a binary characteristic (response) measured successively at two time points for two independent groups. To establish notation, the k-th table (k=1, 2) is of the form

	Response (Y_k)		
Group (X)	1	2	
1	$n_{11.k}$	$n_1 - n_{11.k}$	n_1
2	$n_{21.k}$	$n_2 - n_{21.k}$	n_2

In the table above, $n_{ij.k}$ represent the cell counts of the k-th (k=1, 2) table for group i (i=1, 2) and category response j (j=1, 2). Since we have two independent groups we assume that $n_{11.k}$ and $n_{21.k}$ are independently binomial

distributed, i.e. $n_{11.k} \sim Bin(n_1, \pi_{11.k})$ and $n_{21.k} \sim Bin(n_2, \pi_{21.k})$. Our objective is to estimate the success probabilities $\pi_{11.k}$ and $\pi_{21.k}$ as well as the odds ratios θ_k, comparing the responses of the two groups at each occasion k.

This setup occurs commonly in many clinical and epidemiological studies, where the main issue is the comparison of two independent groups of subjects, in terms of two correlated binary responses for each of them. For example, this type of tables result as marginal tables in case of repeated binary measurements on the same subjects (e.g. before-after treatment) or of simultaneous measurements on the same subjects (e.g. recording of two side-effects).

Here we focus on the case where the individual information, that is (X, Y_1, Y_2) for each subject, is not available. Having observed only the marginal tables of the above type and the corresponding frequencies $n_{ij.k}$, two features need special treatment; the correlation between the two crosstabulations and the non-availability of the complete records for each subject. Such data often arise in practice mainly for reasons of confidentiality protection or due to storage management. Problems of full data availability are also common when the data sources are reports of Governmental Institutes (official statistics). There is no doubt that for categorical data, the traditional and most common form of reporting has been that of marginal tables (Fienberg and Slavkovic, 2004).

If the individual responses are known, many models have been developed in the literature to analyze repeated measurements in the framework of longitudinal studies. Most of them model the marginal expectation of the binary responses using the generalized estimating equations (GEE) methodology (Liang and Zeger, 1986). For example, Fitzmaurice and Laird (1993) modelled time-dependent multivariate binary data by a marginal logistic model. Bayesian models using MCMC algorithms for the analysis of longitudinal data are recommended by Chib and Carlin (1999). In a different framework, Agresti and Klingenberg (2005) proposed a multivariate test comparing the binomial probabilities of two groups in a safety study where each subject was examined in 11 adverse events. However, the methods mentioned so far require the availability of the joint distribution of the multivariate responses for each group.

When analyzing several independent 2×2 tables the Mantel-Haenszel test is often employed. However this approach is not valid in case of intraclass and/or interclass correlation. Most of the modifications of the Mantel-Haenszel test deal with the intraclass correlation. An exception is a procedure given by Begg (1999), which accounts for dependence within and between strata. However, her variance correction factor is based on individual subject information, which is unavailable in the present context (our second special feature). At the same time, Liao (1999) proposed a hierarchical Bayesian model for multiple 2×2 tables, which allows the tables to borrow information from each other. Liao's model is not full Bayesian, since the nuisance parameters are eliminated by conditioning instead of integration.

The estimation of the probabilities $\pi_{11.1}$ and $\pi_{21.1}$ is straightforward, using the cell frequencies of the first table. However, the estimation of $\pi_{11.2}$ and $\pi_{21.2}$ need to take into consideration not only the frequencies of the second table but the correlation between the probabilities $\pi_{i1.1}$ and $\pi_{i1.2}$ (i=1, 2) as well.

Let w_{ij} be the conditional probability of a subject to remain in the same response category at the second measurement, i.e.

$$w_{ij} = P(Y_2 = j | X = i,\ Y_1 = j).$$

Then, the probabilities for the second table are given by $\mathbf{\Pi_2} = \mathbf{W} \cdot \mathbf{\Pi_1}$, where

$$\mathbf{\Pi}_k = (\pi_{11.k}, \pi_{12.k}, \pi_{21.k}, \pi_{22.k})', \quad k = 1, 2$$

and

$$\mathbf{W} = \begin{bmatrix} w_{11} & 1 - w_{12} & 0 & 0 \\ 1 - w_{11} & w_{12} & 0 & 0 \\ 0 & 0 & w_{21} & 1 - w_{22} \\ 0 & 0 & 1 - w_{21} & w_{22} \end{bmatrix}.$$

Note that $\pi_{i2.k} = 1 - \pi_{i1.k}$ for $i = 1, 2$ and $k = 1, 2$.

The Bayesian approach is a natural and convenient choice to deal with missing information. In order to estimate the conditional probabilities w_{ij} we will impose sensible priors to our parameters, which will be non-informative when no prior information is available but can also incorporate prior information otherwise.

49.2 Estimation of Cell Probabilities and Odds Ratios

Let z_{ij} be the unknown number of subjects from the i-th group ($i = 1, 2$) that remained in the j-th response category ($j = 1, 2$) at the second measurement. Since we consider two independent groups, we will describe in detail the analysis for the first one. The results for the second group will be analogous. Thus, the table with the unknown transition frequencies for the first group will be

	$k = 2$		
$k = 1$	$Y_2 = 1$	$Y_2 = 2$	
$Y_1 = 1$	z_{11}	$n_{11.1} - z_{11}$	$n_{11.1}$
$Y_1 = 2$	$n_1 - n_{11.1} - z_{12}$	z_{12}	$n_1 - n_{11.1}$
	$n_{11.2}$	$n_1 - n_{11.2}$	n_1

Given the cell frequency $n_{11.1}$, z_{11} is a binomial observation with parameters $n_{11.1}$ and w_{11}. Similar, given the cell frequency $n_{12.1}(= n_1 - n_{11.1})$, z_{12} is a binomial observation with parameters $n_1 - n_{11.1}$ and w_{12}. Under this consideration,

the full likelihood of the data is

$$L(n_{11.1}, n_{12.1}, z_{11}, z_{12}) \propto \pi_{11.1}^{n_{11.1}} \cdot (1-\pi_{11.1})^{n_1-n_{11.1}} \cdot \frac{w_{11}^{z_{11}} \cdot (1-w_{11})^{n_{11.1}-z_{11}}}{z_{11}! \cdot (n_{11.1}-z_{11})!} \cdot \frac{w_{12}^{z_{12}} \cdot (1-w_{12})^{n_1-n_{11.1}-z_{12}}}{z_{12}! \cdot (n_1-n_{11.1}-z_{12})!} \cdot I(z_{11}^{\min} \leq z_{11} \leq z_{11}^{\max})$$

where $z_{12} = z_{11} + n_1 - n_{11.1} - n_{11.2}$, $z_{11}^{\min} = \max\{0,\ n_{11.1} + n_{11.2} - n_1\}$ and $z_{11}^{\max} = \min\{n_{11.1},\ n_{11.2}\}$.

The priors imposed on the parameters $\boldsymbol{\vartheta}_1 = (\pi_{11.1}, w_{11}, w_{12})$ of interest are independent beta, the usual choice for binomial probabilities

$$\begin{aligned} \pi_{11.1} &\sim \mathcal{B}eta(a, b), \\ w_{11} &\sim \mathcal{B}eta(a_1, b_1), \\ w_{12} &\sim \mathcal{B}eta(a_2, b_2). \end{aligned}$$

When no information is available we set $a_1 = b_1 = a_2 = b_2 = 0.5$, to represent prior ignorance.

In this context, the joined posterior distributions is given by

$$f(\boldsymbol{\vartheta}_1|\boldsymbol{n}_{11}) = f(\pi_{11}|n_{11.1}) f(\mathbf{w_1}|\boldsymbol{n}_{11}),$$

where $\boldsymbol{n}_{11} = (n_{11.1}, n_{11.2})$ and $\mathbf{w}_1 = (w_{11}, w_{12})$. The marginal posterior distribution of π_{11} is a simple beta distribution

$$f(\pi_{11}|n_{11.1}) = \mathcal{B}eta(n_{11.1} + a, n_1 - n_{11.1} + b)$$

while the posterior distribution of $\mathbf{w}_1$ is given by

$$f(\mathbf{w}_1|\boldsymbol{n}_{11}) = \sum_{z_{11}=z_{11}^{\min}}^{z_{11}^{\max}} f_1(w_{11}|z_{11}, n_{11.1}) f_2(w_{12}|z_{11}, \boldsymbol{n}_{11}) f_z(z_{11}|\boldsymbol{n}_{11}) \ ,$$

where the conditional posteriors $f_1(w_{11}|z_{11}, n_{11.1})$ and $f_2(w_{12}|z_{11}, \boldsymbol{n}_{11})$ are beta density functions. The posterior $f(\mathbf{w}_1|\boldsymbol{n}_{11})$ is estimated using a simple MCMC scheme. Note that the conditional posterior $f_z(z_{11}|\mathbf{w_1}, \boldsymbol{n}_{11})$ is a non-central hypergeometric distribution.

Although the resulting posteriors $f_i(w_{1i}|\boldsymbol{n}_{11})$ $(i = 1, 2)$ are not of a recognizable form, it is possible to derive their moments in closed-form expressions. Thus, for the first moments we proved that

$$\begin{aligned} E(w_{11}|\boldsymbol{n}_{11}) &= E(w_{11}|n_{11.1}) = E\left\{E\Big(w_{11}|z_{11}, n_{11.1}\Big)\right\} = \frac{E(z_{11}) + a_1}{n_{11.1} + a_1 + b_1} \ , \\ E(w_{12}|\boldsymbol{n}_{11}) &= E\left\{E\Big(w_{12}|z_{11}, \boldsymbol{n}_{11}\Big)\right\} = \frac{E(z_{11}) + n_1 - n_{11.1} - n_{11.2} + a_2}{n_1 - n_{11.1} - n_{11.2} + a_2 + b_2} \ . \end{aligned}$$

Having estimates of the posterior distributions $f_i(w_{1i}|\boldsymbol{n}_{11})$, $i = 1, 2$, as well as the above expected means, it is possible to estimate the posterior distribution of the success probability $\pi_{11.2}$. The parameter vector for the second group $\boldsymbol{\vartheta}_2$ is defined analogously and the posterior distribution of the corresponding success probability $\pi_{21.2}$ can be estimated accordingly. Finally, the posterior distribution of the odds ratio θ_2 of the second table is derived.

49.3 Example

For illustrative purposes, we analyze a subset of a longitudinal study on the health effects of air pollution carried out at six cities (Ware *et al.*, 1984). This example is very popular in the literature and has been studied under various models for longitudinal data. The data set contains complete records on 537 children from Ohio, each of whom was examined at ages 7 through 10. We will use the summary data for the first and last measurement at ages 7 and 10. The binary response is the wheezing status (1=yes, 0=no) of a child. Maternal smoking (1=if mother was a regular smoker, 0=else) is treated as a fixed variable and forms the two independent groups of children.

49.4 Discussion

In the analysis described so far we assumed that the group sizes (n_1 and n_2) remained fixed for $k = 1$ and $k = 2$, i.e. we assumed that measurements were available on all subjects on both occasions. In case of missing data, if $n_{i.k}$ $(i, k = 1, 2)$ denotes the size of group i at occasion k, then $n_{1.1} \neq n_{1.2}$ and/or $n_{2.1} \neq n_{2.2}$. This context is also under consideration.

It is straightforward to extend our procedure to the case of a response variable with more than two categories. It is also possible to consider the case of analysis of more than two correlated tables ($k > 2$). Finally, we intend to proceed to Bayesian hypothesis testing comparing the probabilities and the odds ratios of the correlated tables.

References

1. Agresti, A. and Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs, *Applied Statistics*, **54**, 691-706.

2. Begg, M. D. (1999). Analyzing k (2×2) Tables Under Cluster Sampling, *Biometrics*, **55**, 302-307.

3. Chib, S.. and Carlin, B. (1999). On MCMC sampling in hierarchical longitudinal models, *Statistics and Computing*, **9**, 17-26.

4. Fienberg, S.E. and Slavkovic, A.B. (2004). Making the release of confidential data from multi-way tables count, *Chance*, **17**, 5-10.

5. Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analyzing longitudinal binary responses, *Biometrika*, **80**, 141-151.

6. Liang, K.Y. and Zeger, S.L (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.

7. Liao, J.G. (1999). A Hierarchical Bayesian Model for Combining Multiple 2×2 Tables Using Conditional Likelihoods, *Biometrics*, **55**, 268-272.

8. Ware, J.H., Dockery, D.W., Spiro, A.III, Speizer, F.E. and Ferris, B.G. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities, *Am. rev. Respir. Dis.*, **129**, 366-374.

50

Latent Class Analysis to Evaluate the Accuracy of Diagnostic Tests for Leishmaniasis

Filipa Encarnação†, Luzia Gonçalves‡, Lenea Campino‡, José M. Cristovão‡ and M. Rosário de Oliveira†

† *Dep. de Matemática and CEMAT, Instituto Superior Técnico, Av. Rovisco Pais, 1049 - 001 Lisboa, Portugal*
afilipa.vieira@gmail.com, rosario.oliveira@math.ist.utl.pt

‡ *Instituto de Higiene e Medicina Tropical - UNL, Rua da Junqueira, 96, 1349-008 Lisboa, Portugal*
{*LuziaG,campino*}*@ihmt.unl.pt*

Abstract:

The latent class model is a common alternative to evaluate the accuracy of diagnostic tests in the absence of a true gold standard. In this paper, we adjusted a latent class model with constraints to a data set of stray dogs in order to estimate sensitivities and specificities of four diagnostic tests for *Leishmania* infection. The constraints incorporate certain properties of the parasitological tests. In the adjusted model, the values of sensitivity and specificity for the parasitological test (applied to liver and spleen) and bone marrow PCR are the same. In practice, spleen and liver biopsies can only be done in dead dogs. Consequently, we recommend bone marrow PCR as a good test to detect this infection.

Keywords and phrases: Latent class analysis, Sensitivity, Specificity, Leishmaniasis

50.1 Introduction

Leishmaniasis is a group of diseases caused by parasites of the genus *Leishmania* (Campino, 2002). These parasites transmitted by sand fly can infect a variety of hosts, including humans and dogs. Farrell (2002) focuses that "*human leishmaniasis occurs in tropical, sub-tropical and temperate regions of the world with an estimated 1.5 to 2 million new cases each year*". The dog is an important reservoir that serves as the source of human infection (Campino, 2002).

In this work, we considered a set of $n = 132$ stray dogs collected in the outskirts of Lisbon. A dog can be infected without showing any clinical signs of the disease. Consequently, clinical diagnosis must be confirmed or assessed

through laboratory tests. Serologic tests (indirect immunofluorescence, IFI, and counterimmunoelectrophoresis, CIE), parasitological tests (parasite detection in traditionally affected tissues through microscopy and cultures, Par*Tissue*) and DNA tests (Polymerase Chain Reaction, PCR) are common in the laboratory diagnosis of leishmaniasis in dogs (and humans), but the results obtained are not confirmatory.

In medical applications, diagnostic tests give indications of whether or not an individual has a certain disease (or infection). Consequently, the study of their performance is a common problem of major importance. Suppose that p diagnostic tests are applied independently to each subject and let X_i ($i = 1, \ldots, p$) be the result of the i-th diagnostic test, taking the value 1 if the subject is diagnosed as having the disease (positive result) and 0 (negative result) otherwise. For each subject, the true state of the disease is a variable, Y, that can take one of two values: 1 indicates that the subject has the disease, and 0 that the subject does not have the disease. Sensitivity and specificity are measures of the performance of a diagnostic test. Sensitivity is the probability of a diseased individual to be correctly identified by the test, $P(X_i = 1|Y = 1)$, and specificity is the probability of a healthy individual being correctly identified by the same test, $P(X_i = 0|Y = 0)$.

Usually, these measures are calculated comparing tests with a reference test known as "gold standard". In an ideal situation, this gold standard is known to be capable of correctly classifying an individual as diseased or non-diseased (infected or non-infected). However, in practice, gold standards are rare and due to this, the true state of the disease cannot be assessed. Thus, it can be seen as a latent variable, and the sensitivity and specificity of the diagnostic tests being studied can be estimated using latent class analysis (Bartholomew and Knott, 1999; Hadgu and Qu, 1998).

In Section 50.2 we introduce the latent class model with constraints that express certain properties of the parasitological tests. In Section 50.3 the main results as well as the validation of the models are discussed. The paper ends with some conclusions.

50.2 Latent Class Model

Let $\eta_1 = P(Y = 1)$ be the prevalence of the disease (or infection) and $\pi_{ij} = P(X_i = 1|Y = j)$, $i = 1, \ldots, p$ and $j = 0, 1$. Using this notation, the sensitivity and the specificity of the i-th test can be written as π_{i1} and $1-\pi_{i0}$, respectively.

The latent class model assumes that, given the true state of the disease, the results of the diagnostic tests are independent. This assumption is called hypothesis of conditional independence (HCI) and in some medical problems may not be a realistic assumption, which is the reason it has to be validated. In this work, goodness-of-fit tests as well as the correlation residual plot suggested

by Qu *et al.* (1996) are performed to validate this hypothesis.

It is known that if a parasite is detected in a tissue then the subject is infected. However, if a parasite is not detected, we cannot conclude that the subject is not infected. Let X_k ($k = 1, \ldots, q$ and $q < p$) be the result of the k-th parasitological test applied to a certain type of tissue. Thus, if the parasite is detected in the k-th tissue, $X_k = 1$, then the subject is infected, $Y = 1$, and due to this we can write: $P(Y = 1|X_k = 1) = 1$. This is equivalent to stating that the specificity of the k-th parasitological test is equal to 1, i.e. $\pi_{k0} = 0$, $k = 1, \ldots, q$. Incorporating these constraints in the latent class model we can estimate its parameters using the EM-algorithm. Let $\boldsymbol{x}_h = (x_{h1}, \ldots, x_{hp})^t$ be the response vector associated with the h-th subject. Given some initial values, the iterative estimation procedure can be formulated as follows:

$$\hat{d}(1|\boldsymbol{x}_h) = \frac{\hat{\eta}_1 \prod_{i=1}^{p} (\hat{\pi}_{i1})^{x_{hi}} (1 - \hat{\pi}_{i1})^{1-x_{hi}}}{\hat{P}(\boldsymbol{X} = \boldsymbol{x}_h)}$$

where

$$\begin{aligned}\hat{P}(\boldsymbol{X} = \boldsymbol{x}_h) &= \hat{\eta}_1 \prod_{i=1}^{p} (\hat{\pi}_{i1})^{x_{hi}} (1 - \hat{\pi}_{i1})^{1-x_{hi}} + \\ &\quad + (1 - \hat{\eta}_1)\delta(\sum_{k=1}^{q} x_{hk}) \prod_{i=q+1}^{p} (\hat{\pi}_{i0})^{x_{hi}} (1 - \hat{\pi}_{i0})^{1-x_{hi}}\end{aligned}$$

and $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. Note that $\hat{d}(0|\boldsymbol{x}_h) = 1 - \hat{d}(1|\boldsymbol{x}_h)$, and

$$\hat{\eta}_j = \frac{1}{n} \sum_{h=1}^{n} \hat{d}(j|\boldsymbol{x}_h), \qquad \hat{\pi}_{ij} = \frac{\sum_{h=1}^{n} \hat{d}(j|\boldsymbol{x}_h) x_{hi}}{n\,\hat{\eta}_j},$$

$i = 1, \ldots, p \quad j = 0, 1.$

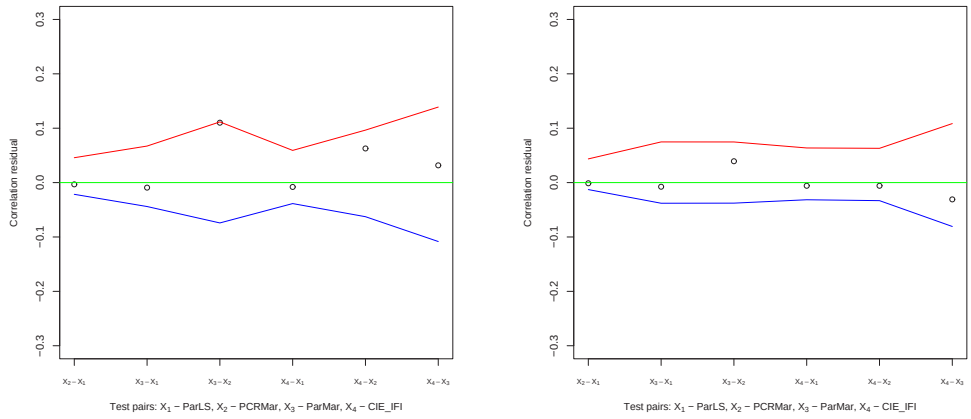
50.3 Results and Discussion

Several diagnostic tests were applied independently to each dog, but only $p = 4$ variables were considered in this study. The two serological tests (IFI and CIE) were used to build a new variable, named CIE_IFI, such that CIE_IFI=CIE*IFI.

$\hat{\eta}_1$			ParLS	PCRMar	ParMar	CIE_IFI
$n = 132$	0.243	Specificity	1.000	1.000	1.000	0.960
		Sensitivity	0.968	0.906	0.781	0.813
$n = 130$	0.231	Specificity	1.000	1.000	1.000	0.960
		Sensitivity	0.966	0.966	0.833	0.867

Table 50.1: Estimates of the prevalence, sensitivities and specificities when we consider $n = 132$ (all data set) and $n = 130$, when we exclude two problematic dogs from the analysis.

This means that CIE_IFI is positive only if both CIE and IFI are positive. These variables were combined in order to overcome the violation of HCI.



(a) $n = 132$. (b) $n = 130$.

Figure 50.1: Correlation residual plot with a 95% bootstrap confidence band, estimated using a percentile method (Davison and Hinkley, 1997).

Three parasitological tests were performed on different tissues: bone marrow (ParMar), liver (ParL) and spleen (ParS). Since the liver and spleen were not our major concern, a new parasitological variable was built, ParLS, such that whenever a parasite is detected in the liver or in the spleen then ParLS=1, and ParLS=0 only if no parasite is detected in either of these two tissues. The other variable under study was the PCRMar, the PCR done on the bone marrow of each dog. The results of the analysis on the $n = 132$ dogs are summarized in Table 50.1. In order to validate these results we considered two goodness-of-fit tests: likelihood ratio test (p-value: 0.023) and a bootstrap test (p-value: 0.002). The tests indicate that we should reject the adequacy of the model to the data. Figure 50.1 (a) shows that the correlation residual between ParMar and PCRMar touches the 95% bootstrap confidence band (Davison and Hinkley, 1997), indicating violation of the HCI. Looking at the data more carefully, we

conclude that the model is rejected mainly because of two dogs. In both cases, only ParS detected the parasite and all the other tests (ParMar, PCRMar, CIE_IFI) gave negative results. According to the latent class model, a response vector like this would be very unlikely, and an observed frequency of 2 is quite large when compared with the corresponding expected frequency, 0.12.

In order to overcome this difficulty we eliminated the two problematic dogs from the data set and repeated the latent class analysis. The new results are summarized in Table 50.1 ($n = 130$). In this case, all the tests accepted the adequacy of the model to the data (the p-values for the log-likelihood ratio and bootstrap tests are 0.533, 0.173, respectively) and the Qu *et al.* (1996) correlation residual plot gives no indication of the violation of the HCI (see Figure 50.1 (b)). When we exclude the two dogs the specificities are unchanged, however, the estimated sensitivities of ParMar, PCRMar, CIE_IFI get higher and the estimated sensitivity of ParLS gets slightly lower. This was expected since we removed two infected dogs from the analysis that ParMar, PCRMar and CIE_IFI have not detected. Note that the new estimated prevalence is lower, which was also expected.

50.4 Conclusions

Dogs can be infected without showing any clinical signs of the disease. Consequently, their clinical diagnosis has to be assessed through diagnostic laboratory tests. In order to evaluate the accuracy of diagnostic tests for *Leishmaniasis* we presented the latent class model with constraints that express certain properties of the parasitological tests (specificity equal to 1).

We studied a data set of 132 stray dogs, but all the models were rejected by the goodness-of-fit tests and by the correlation residual plot. The main reason seems to be the two problematic dogs. Repeating the whole analysis without the two correspondent response vectors we did not find any evidence to reject the latent class model. Through this we may conclude that the best diagnostic tests are bone marrow PCR (PCRMar) and the parasitological test that combines results obtained from liver and spleen (ParLS). Despite having the same values of sensitivity and specificity (1.00 and 0.966, respectively), we choose PCRMar as the recommended diagnostic test because, in practice, it is more feasible than ParLS, since spleen and liver biopsies can only be done in dead dogs.

In addition, we conclude that the prevalence is approximately $23 - 24\%$. This may be explained by the fact that all the analysed dogs are stray, living under bad hygiene and health conditions.

References

1. Bartholomew, D.J. and Knott M. (1999). *Latent Variable Models and Factor Analysis*. (2nd Edition). London: Arnold.

2. Campino, L.M. (2002). Canine Reservoirs and Leishmaniasis: Epidemiology and Disease, In *Leishmania*. (Ed., Jay P. Farrell), pp. 45–57, Kluwer Academic Publishers.

3. Davison, A. C. and Hinkley, D. V.(1997). *Bootstrap Methods and their Application*. Cambridge University Press.

4. Farrell, J.P. (2002). *Leishmania*. (Ed., Jay P. Farrell). Kluwer Academic Publishers.

5. Hadgu A. and Qu Y. (1998). A Biomedical Application of Latent Class Models with Random Effects, *Appl. Statist.*, **47**, 603–616.

6. Qu Y., Tan M. and Kutner M. (1996). Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Test, *Biometrics*, **52**, 797–810.

51

Applying Functional Data Models to Predict the Burden of Breast Cancer in USA and UK

Bircan Erbas[1], Rob Hyndman[2], Muhammad Akram[2] and Dorota Gertig[1]

[1] *Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne*
[2] *Department of Econometrics and Business Statistics, Monash University*

Abstract: Accurate estimates of future age-specific incidence and mortality are critical for allocation of resources to breast cancer control programs and evaluation of screening programs. There has been very little advancement in developing statistical methodological for cancer projections in recent years. The aim of this study is to apply a new approach proposed by Erbas, Hyndman and Gertig (2006) to forecast age-specific breast cancer mortality in the US and UK. This method has potential application as an alternative tool to evaluate the effectiveness of mammographic screening on mortality from breast cancer. Moreover, these models have broader application to other cancers and chronic diseases.

Keywords: Functional Data Models, Breast cancer, forecasting

51.1 Introduction

Despite increased utilization of mammographic screening and continual improvements in treatment options, breast cancer remains one of the main causes of mortality and morbidity in women. Each year millions of dollars are spent on development of effective strategies for cancer control programs and planning of services, and accurate estimates of future age-specific incidence and mortality are necessary to support health care organisations in preparing recommendations for allocation of resources to breast cancer control programs. Furthermore, the widespread use of mammographic screening raises important policy questions regarding the mortality benefits of early detection and is a major factor to be considered when estimating future incidence and mortality.

There has been little advancement in developing statistical methodologies for cancer projections in recent years. Variants of age-period-cohort methods have been used to project mortality and incidence from breast cancer (Dyba and

Hakulinen (2000); Blanks *et al.* (2000); Bashir and Esteve (2001); Moller *et al.* (2002)). However, these methods are inadequate in capturing the shape of the mortality-age relationship as it varies with time and projections are sensitive to the most recent changes in cohort effects.

The objective of this study is to apply functional data analysis techniques to model trends in US and UK breast cancer mortality rates, treating the observed data as curves with age as a functional covariate. These trends will be used to forecast the entire age-specific mortality function for future time periods using an exponential smoothing state-space approach for forecasting. This new method allows the shape of the incidence-age curve to vary with time so that, at different ages, mortality declines at different rates, a phenomenon which is particularly apparent for breast cancer. Current methods simply extrapolate most recent trends into the future but our approach forecasts the entire age-mortality function with features that will likely increase forecast accuracy.

51.2 Methods

We use functional data analysis techniques (Ramsay and Silverman, 2005) where time trends of mortality are modelled as annual curves with age as a functional covariate. Specifically, let $y_t(x)$ denote the curve over x for time period t. Time is considered to be discrete and x is assumed to be continuous. However, usually the curves are observed for discrete values of x. We assume there is an underlying smooth function $f_t(x)$ that we are observing with error. Thus, we observe the functional time series $\{x_i, y_t(x_i)\}$, $t = 1, \ldots, n$, $i = 1, \ldots, p$ where

$$y_t(x_i) = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}\,, \tag{51.2.1}$$

$\varepsilon_{t,i}$ is an iid random variable with zero mean and unit variance and $\sigma_t(x_i)$ allows the amount of noise to vary with x. In cancer epidemiology $\{x_1, \ldots, x_p\}$ usually denote 5-year age groups. The smooth curves $\{f_t(x)\}$ can be estimated using nonparametric regression techniques such as regression splines.

Erbas, Hyndman and Gertig (2006) and Hyndman and Ullah (2005) proposed the following functional time series model

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x) \tag{51.2.2}$$

where $\mu(x)$ is a measure of location of $f_t(x)$, $\{\phi_k(x)\}$ is a set of orthonormal basis functions, and $\beta_{t,k}$ is a univariate time series. They computed $\{\phi_k(x)\}$ using functional principal components decomposition (Ramsay and Dalzell, 1991) applied to the smooth curves $\{f_t(x)\}$ as this approach gives a small number of basis functions, enables informative interpretations and gives coefficients which are uncorrelated with each other.

The coefficients $\{\beta_{t,k}\}$ are each forecast using a univariate time series model, thus giving forecasts of $f_{n+h}(x)$, $h = 1, 2, \ldots$. We use exponential smoothing state space models due to Hyndman et al. (2002).

The models are applied to US and UK age-specific breast cancer mortality data. We obtained US breast cancer mortality data (1950–2001) from the National Cancer Institute and UK (including Wales) data (1950–2003) from the Office for National Statistics. We use crude (unadjusted) age-specific mortality rates from breast cancer expressed as per 100,000 people.

51.3 Results

51.3.1 US age-specific breast cancer mortality

A functional regression model with $K = 2$ basis functions accounts for 92.5% of the variation around the mean mortality curve. (The first two basis functions explain 68.4% and 24.1% of the proportion of variation respectively.) A set of $K = 4$ basis function minimized the mean integrated square error (MISE). Diagnostics of the model with two basis functions show an adequate fit of the data.

Ten year forecasts of the first basis function for US all-race breast cancer mortality and corresponding 80% prediction intervals are presented in Figure 51.1. Preliminary analysis suggests overall breast cancer mortality rates in the US will continue to decline. Figure 51.2 shows age-specific mortality projections. The early years are represented as red, orange, yellow, green, blue with violet

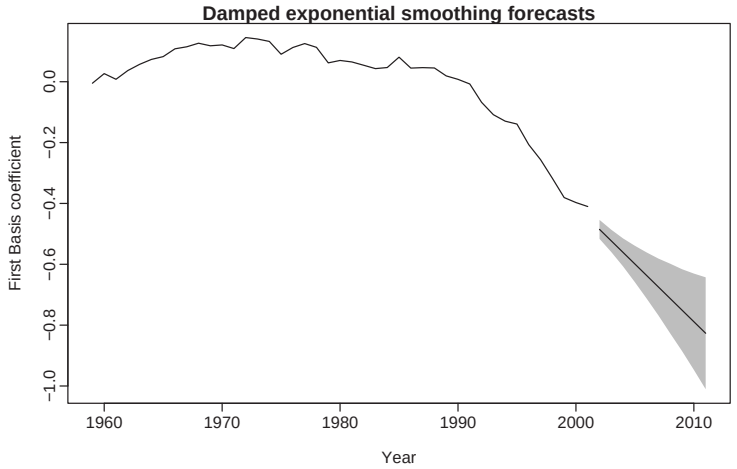


Figure 51.1: *Ten year forecasts of the first coefficient using an exponential smoothing model for US age-specific breast cancer mortality. The shaded region gives 80% prediction intervals.*

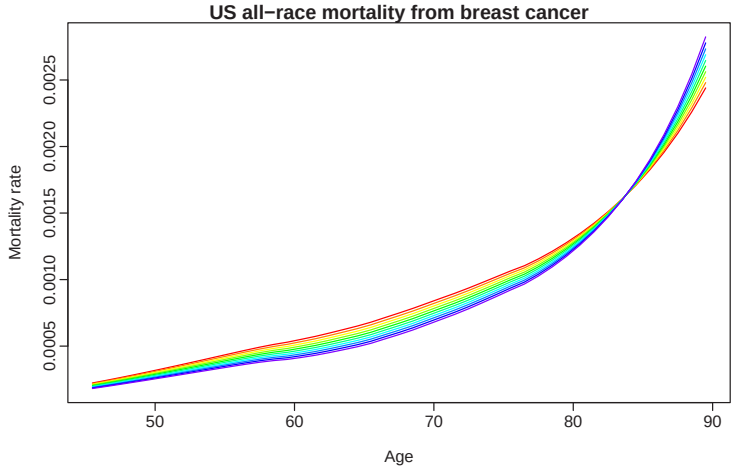


Figure 51.2: *US age-specific mortality from breast cancer.*

representing the most recent year. Mortality rates for women less than 80 years of age will continue to decline. However, mortality rates are expected to slightly increase for women between 80 and 90 years of age.

51.3.2 UK age-specific breast cancer mortality

Preliminary analysis suggests a functional time series model with $K = 3$ basis functions explain 96.8% of variation around the mean mortality curve. (The first three basis functions represent 71.6%, 21.4% and 3.8% of total proportion of variation respectively.)

Figure 51.3 displays ten year forecasts of UK breast cancer mortality rates using a damped trend exponential smoothing forecasting model. These predictions suggest a continual decline in breast cancer mortality rates but at a slower rate compared to recent years. The age-specific 10 year forecasts are displayed in Figure 51.4. Mortality rates are expected to continue to decline (slightly) for women between 50 and 75 years of age, while mortality is expected to remain constant for women over 80 years of age.

51.4 Discussion

Future predictions of mortality and incidence from breast cancer aid public health administrators in formulating policy on allocation of resources to screening or treatment in different age groups of women. This approach, proposed by Erbas, Hyndman and Gertig (2006), has never been applied to study age-specific mortality of breast cancer in the US and UK. In this study we demonstrate the utility of these methods in further understanding the behaviour of age-specific

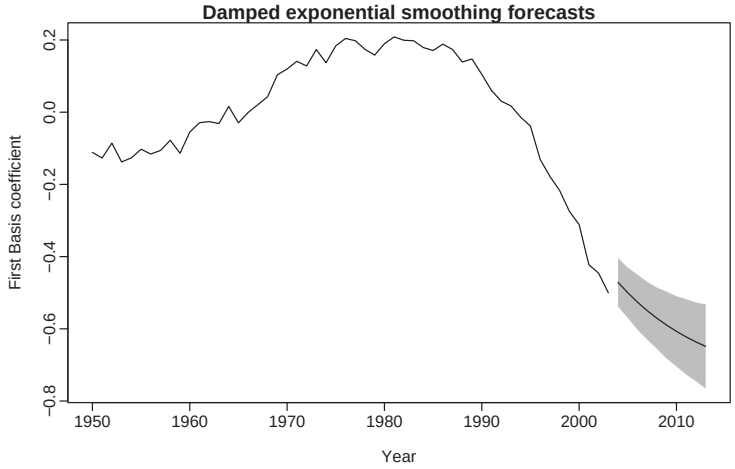


Figure 51.3: *Ten year forecasts of UK age-specific breast cancer mortality for the first coefficient using an damped trend exponential smoothing model. The shaded region gives 80% prediction intervals.*

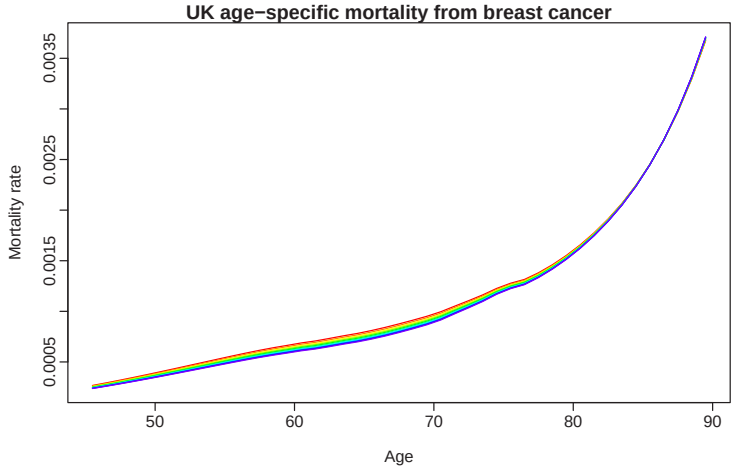


Figure 51.4: *UK age-specific mortality from breast cancer.*

mortality trends of both US and UK data, and estimate future predictions of the breast cancer burden on these populations.

As mammographic screening is the main form of early detection, it is important to assess the impact of screening on mortality and incidence and incorporate the effects of screening in future predictions. With the large investment in mammographic screening, tools that can reliably provide accurate estimates of future predictions of incidence and mortality in the presence of screening are essential. These models have the potential to incorporate both screening and

treatment effects. Furthermore, in our future work we will extend these models to incorporate cohort effects, important components of the overall age-specific mortality and incidence trends.

References

1. Dyba T, and Hakulinen T (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques, *Statistics in Medicine*, **19**, 1741–1752.

2. Blanks RG, Moss SM, McGahan CE, Quinn MJ, and Babb PJ (2000). Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990-8: comparison of observed with predicted mortality, *British Medical Journal*, **321**, 665–669.

3. Bashir SA, and Esteve J (2001). Projecting cancer incidence and mortality using Bayesian age-period-cohort models, *Journal of Epidemiology and Biostatistics*, **6**, 287–296.

4. Moller B, *et al.* (2002). Predictions of cancer incidence in the Nordic countries up to the year 2020, *European Journal of Cancer Prevention*, **11**, S1–S96.

5. Ramsay JO and Silverman BW (2005) *Functional data analysis*, Springer-Verlag: New York.

6. Erbas B, Hyndman RJ and Gertig D (2006). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, to appear.

7. Hyndman RJ, and Ullah S (2005) Robust forecasting of mortality and fertility rates: a functional data approach, Department of Econometrics and Business Statistics working paper. Available at <`http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2005/wp2-05.pdf`>

8. Ramsay JO, and Dalzell CJ (1991). Some tools for functional data analysis with discussions, *Journal of the Royal Statistical Society Series B*, **53**, 539–572.

9. Hyndman, R.J., Koehler, A.B., Snyder, R.D., and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, **18(3)**, 439-454.

52

The Effect of Speech Disorders on the Quality of Life

Nicolas Farmakis*, Mavroudis Eleftheriou*, Diomidis Psomopoulos**

** Aristotle University of Thessaloniki, Department of Mathematics, Sector of Statistics and Operational Research*
**** Diagnostic and Treatment Center of Speech and Language Disorders*

Abstract: Considering the social consequences of speech and language disorders, the demand of a statistical analysis is imperative. In this analysis, basic statistics for the types of speech and language disorders, for their frequency in the two genders, for the duration of the treatment and for the results of the treating procedure are derived and discussed.

Keywords and phrases: Speech and language disorders, Quality of life, Statistical analysis

52.1 Introduction

Speech and language disorders are defined as the disabilities (inborn or acquired) of children to use their mother tongue in the considered normal rules of their development. It should be noted that speech and language disorders can be transient, long running or permanent and can be extended to one or more of the fundamental levels of speech.

The quality of life relies greatly on the ability of speaking correctly. Speech and language disorders can disturb the normal development of children and affect their emotional and psychological status as well as their social adaption.

Worldwide researches, claim that speech and language disorders emerging among children at preschool age, reach 25%. This percentage decreases to 10% when children enter school and constitutes the basic set of the children needing systematic treatment.

The statistical analysis of data related to speech and language disorders helps to plan effectively the treating of these disorders. The following study

attempt to serve this purpose, analyzing a collection of observations relevant to the ability of children of age at most 16, to speak correctly.

52.2 Description of data

The data is derived from a diagnostic and treatment center located in Thessaloniki, Greece, treating children with speech and language disorders, from 1999 until 2005. The statistically significant size of the sample and the wide chronic spectrum of the observations, render the conclusions efficient and reliable (adhering to the findings of the most recent relevant studies and researches). The data are organized in four variables, fundamental for the analysis:

Disorder: This variable describes the type of speech and language disorder. The most frequent disorders have been included in the analysis and have been categorized into eight groups: *Learning disorders, Articulation disabilities, Developmental delay, Autism, Stuttering, Hearing disorders, Cerebral palsy, Rare syndromes.*

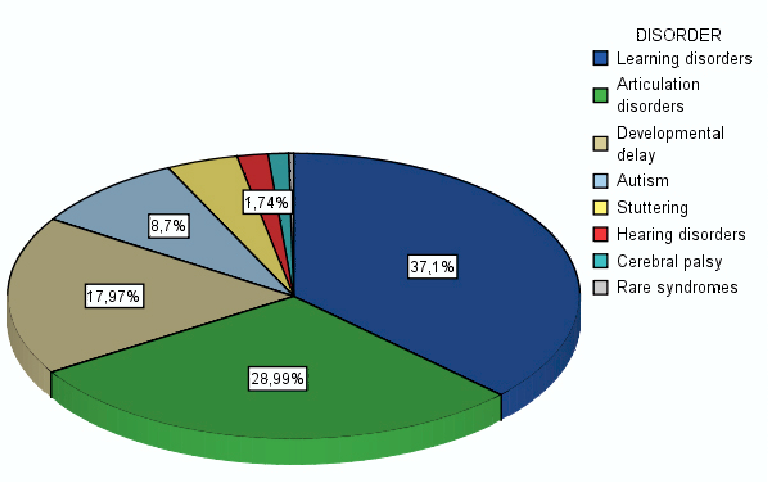
Sex: This variable indicates the gender of each participant.

Therapy duration: The duration of the therapy until its completion, implying the achievement of the maximum degree of improvement, until the day the child is withdrawn from the treating procedure (for reasons irrelevant to this study) or until today, if the treating procedure is still in progress. The validity of the analysis is secured, discarding the observations corresponding to very small periods of therapy (less than 25 days) in which the determination of a child's problem and the evaluation of its progress is not feasible.

Outcome: The outcome of the therapy until its completion, the withdrawal of the participant or until today if the therapy is still in progress. There are three levels indicating the improvement of each participant: slight improvement, satisfying improvement and significant improvement.

52.3 Descriptive statistics

The following pie graph indicates the allocation of the observations amongst the 8 basic disorders.



It is derived that 70,1% of participants are males and only 29,9% females. The significant discrepancy of the percentages of the two sexes is predictable. Various statistical researches confirm that the ratio of males and females with speech and language disorders is approximately 7/3. Thus, the above percentages are indicative of the validity of the data. It is also obvious that 45,2% of the participants are currently in therapy. The percentage of participants who conclude successfully the therapy is approximately equal to that of participants who quitted the therapy before its completion.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Success	96	27,8	27,8	27,8
	In progess Therapy	156	45,2	45,2	73,0
	Stopped	93	27,0	27,0	100,0
	Total	345	100,0	100,0	

It is derived that the mean value of therapy duration significantly differs not only among the two sexes but also among the eight basic categories of speech and language disorders. However it is easily verified that due to a large number of outliers, a most representative estimation of the mean value would be offered by the M-estimators. All of them suggest a mean value for males around 300 and for females close to 500 days.

It is interesting to test statistically the equality of mean values of therapy duration in the categories of sex and disorder. Considering the populations of the two sexes as independent, a t-test was conducted and rejected the hypothesis of equal mean values. For the disorder variable, our interest was focused on the most frequent disorders, learning disorders and articulation disabilities. The t-test verifies that the mean values of therapy duration in these two categories, differ significantly. It must be noted that the use of t-test is validated due to the large number of participants.

Considering that a significant percentage of the participants (45,2%) correspond to unfinished treatment, it is of major priority to attempt one more exploration of the data discarding them. The t-test for one more time indicates the significant difference of the mean value of therapy duration in the categories both of sex and disorder. It is also derived that the mean values of therapy duration differ in the categories of outcome for the same selection of data, implying that the duration of therapy straightforwardly affects its final outcome.

Aiming at a comprehensive exploration of the data, we conducted a two-way independent ANOVA procedure, attempting to reveal the significance of the factors and their interactions. Firstly, the following table of all coefficients and their interactions is presented. Obviously, the only significantly unequal to zero coefficients are sex, disorder and the interaction of outcome and disorder. This model would explain the 25,2% of the data variability. This percentage is not statistically significant but it is indicative of the role of the type of disorder and of gender in the configuration of therapy duration.

Tests of Between-Subjects Effects

Dependent Variable: THERAPY_DURATION

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	17276342,9[a]	33	523525,541	3,161	,000
Intercept	21639312,4	1	21639312,42	130,666	,000
OUTCOME	989874,144	2	494937,072	2,989	,052
SEX	676937,916	1	676937,916	4,088	,044
DISORDER	3432637,716	6	572106,286	3,455	,003
OUTCOME * SEX	356401,067	2	178200,533	1,076	,342
OUTCOME * DISORDER	3674192,906	11	334017,537	2,017	,026
SEX * DISORDER	739468,689	4	184867,172	1,116	,349
OUTCOME * SEX * DISORDER	313829,618	6	52304,936	,316	,929
Error	51338323,7	310	165607,496		
Total	149977909	344			
Corrected Total	68614666,6	343			

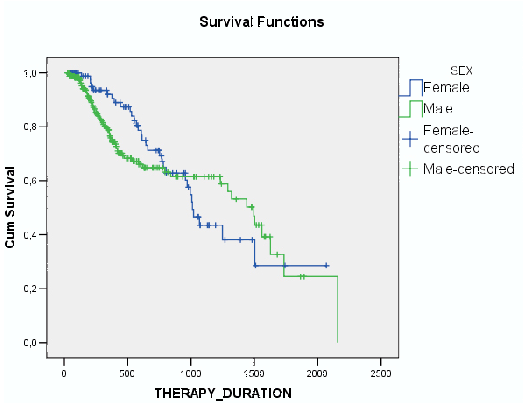
The Post Hoc tests verify that the mean value of therapy duration was different for the observations of slight and significant improvement, as well as for the observations of satisfying and significant improvement. That means that the improvement is greater as longer as the therapy is. This conclusion can also be affirmed by planned comparisons. It is also indicated by the Post Hoc tests that the learning disorders-articulation disabilities and articulation disabilities-developmental delay are pairs with significantly different mean values of therapy duration.

It is also important to examine how the outcome varies by sex or disorder. A crosstabulation procedure leads us to the conclusion that outcome, sex and disorder probably are not significantly dependent for these data. Few of the symmetric directional measures however reject the null hypothesis of independence, for the males. This fact indicates dependence of the disorder and outcome which may not be clear due to bias of the data. Moreover the crosstabulation procedure confirms the dependence of the type of disorder by

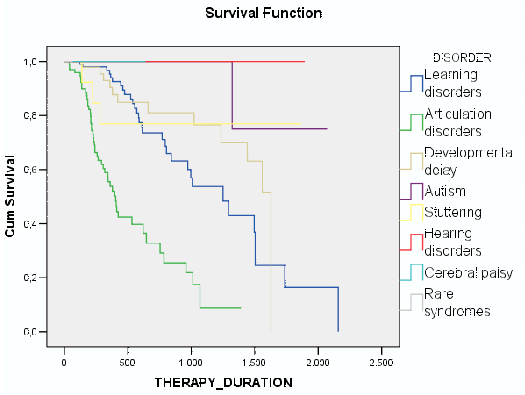
sex. More girls suffer from learning disorders and articulation disabilities while more boys seem to suffer from more serious and harder to treat disorders like developmental delay, autism etc.

52.4 Survival analysis

Considering the therapy duration as survival time and consequently the data as survival data, a Kaplan-Meier model can be constructed. The event of completing successfully the therapy corresponds to a "death" event.



The observations corresponding to therapy in progress and termination of the therapy represent the censored data. Any point of the graph above derived by Kaplan-Meier procedure, shows the probability that a child of a given gender will not have reached full treatment by that time. The plot also suggests that boys may be treated faster than girls when the therapy procedure is less than 1000 days. After this time the girls seem to attain successful results easier than boys. This difference between the two curves is statistically significant as the Breslow (generalized Wilcoxon) test verifies.



The plot above represents the survival curves of all types of speech and language disorder. Any point of it shows the probability that a child suffering from the given disorder will remain in the therapy procedure. Focusing on the two basic disorders, learning disorders and articulation disabilities, it is obvious that children suffering from articulation disabilities achieve treatment faster than the children with learning disorders. Pairwise comparisons lead us to conclude that the survival curves of learning disorders and articulation disabilities are statistically different.

52.5 Conclusions

Although the current analysis is focused on narrow geographical margins, it has the prospect to illuminate and verify the statistics relevant to the area of speech and language disorders. It presents the allocation of speech and language disorders in the two genders, verifying that boys have a stronger tendency to suffer from them. Mean values of therapy duration were derived for the two genders and the basic speech and language disorders and were examined for statistically significant deviance. It has been shown that girls achieve successful treatment faster than boys while articulation disabilities are treated faster than the next most frequent disorder, learning disorders. Obviously the degree of treatment varies by gender and disorder.

References

1. Farmakis N. (2002). Introduction to Sampling, Christodoulidis Publ., Thessaloniki.

2. Mrija J. N. (2002). SPSS 11 Guide to Data Analysis, Prentice Hall.

3. Otto B. (1999). Sprachstoerungen bei kindern und Jugendlichen, Kohlhammer, Stuttgart.

4. Heinz B. (1979). Sonder Paedagogik im Grudriss, Marhold, Berlin.

5. Walpole R. (1974). Introduction to Statistics, 2nd Ed. McMilan Publ. Co., Inc. New York.

6. Kounias S., Kolyva-Mahera F., Bagiatis C., Bora-Senta E. (1985). Introduction to Statistics, Christodoulidis Publ. Co., Thessaloniki.

53

Nested Plans For The Change Point Problem In Exponential Families

Paul D. Feigin, Gregory Gurevich and Yan Lumelskii

Technion - Israel Institute of Technology, Haifa, Israel
Sami Shamoon College of Engineering, Beer Sheva, Israel
Technion - Israel Institute of Technology, Haifa, Israel

Abstract: We consider the problem of sequential quality control and propose nested plans for the early detection, with low false alarm rate, of a change in a stochastic system. The nested plan includes two phases: a variable plan and an attributes plan. For the proposed specific nested plan we present the exact (non-asymptotic) expressions for the mean and the standard deviation of the run length to false alarm and the delay in detection. We assume that the initial and the final distributions come from an exponential family of distributions. The multivariate normal distribution is considered specifically.

Keywords and phrases: Nested plan, change detection, exponential family of distributions, multivariate normal distribution, average run length

53.1 Introduction and Notation

There are extensive references in statistics and engineering on the subject of early detection, with low false alarm rate, of parameter changes in stochastic systems on the basis of sequential observations from the system. Such problems are very important in the context of quality and reliability control(see Lai (1995), Zacks (1991)). In this talk we consider nested plans for the early detection of a parameter change assuming an exponential family situation. We continue research which was started in articles Feigin *et al.* (2005), Lumelskii and Feigin (2005), Lumelskii *et al.* (2006).

Often instead of individual observations $X_1, X_2, \ldots$ one has a sequence of samples of n observations. We assume that the process under investigation yields independent samples, each of n independent observations $(X_{11}, X_{12}, \ldots, X_{1n})$, $(X_{21}, X_{22}, \ldots X_{2n}), \ldots$. Initially these observations follow a distribution $F(x \mid \theta_1)$. At m, an unknown point in time, something happens to the process,

causing the distribution of the sample's observations to change to $F(x \mid \theta_2)$; $F(x \mid \theta_1) \neq F(x \mid \theta_2)$. In this article we assume that the distribution $F(x \mid \theta)$ is of exponential family form.

A common performance measure for any inspection scheme is the average run length (ARL). Let T be the time when the scheme signals that the process is out of control (distribution of the observations has changed). The ARL is defined by $E_{F(x|\theta_1)}T$ where we define $E_{F(x|\theta_h)}T \equiv E(T \mid \theta_h)$ the expectation of the stopping time T under the assumption that the observations come from some distribution $F(x \mid \theta_h)$.

Clearly, one wants $E(T \mid \theta_1)$ to be large and $E(T \mid \theta_2)$ to be small. For the situation with univariate observations there are known optimal CUSUM control charts. Alternatively, Shiryayev-Roberts control charts have also been recommended as nearly optimal. However, the practical design of such charts is not simple because there are no simple explicit expressions for the ARL and $E(T \mid \theta_2)$. As a result, these schemes are geared toward detecting small changes, whereas for large changes Shewhart charts are typically designed.

We propose nested plans, denoted $\prod_{nes}(\Pi^{G_1}|\Pi^{G_2})$, for the quick detection of a change in the distribution of observations. It consists of two steps, Π^{G_1} and Π^{G_2}: Π^{G_1} is a variables plan; whereas Π^{G_2} is an attributes plan (see Feigin *et al.* (2005)).

We consider the fist step of the nested plan with parameters n and C (n is natural number, size of sample; C is real constant). Using sequential observations $(X_{11}, X_{12}, \ldots, X_{1n})$, $(X_{21}, X_{22}, \ldots, X_{2n}), \ldots$ with distribution $F(x \mid \theta_h)$, the first step of the nested plan is given by

$$Y_i(\theta_1, \theta_2) = \ln \prod_{j=1}^{n} \frac{f(X_{ij} \mid \theta_2)}{f(X_{ij} \mid \theta_1)}, \quad i = 1, 2, \ldots \tag{53.1.1}$$

and

$$Z_i = \begin{cases} 1, & if \quad Y_i(\theta_1, \theta_2) > C, \\ 0, & if \quad Y_i(\theta_1, \theta_2) \leq C. \end{cases} \tag{53.1.2}$$

Defining

$$P_h \equiv P(Z_i = 0 \mid \theta_h) \equiv P(Y_i(\theta_1, \theta_2) \leq C \mid \theta_h), \quad Q_h = P(Z_i = 1 \mid \theta_h), \tag{53.1.3}$$

we have a binary sequence of observations Z_1, $Z_2, \ldots$ with probability of zero equal to P_h.

53.2 Second Stage and Characteristics of Nested Plans

The second step is an attributes plan Π^{G_2}, which is based on $Z_i = 0$ or $Z_i = 1$ $i = 1, 2, \ldots$. We consider the plan $\Pi^2(d; 2)$ for which the stopping rule is defined as $T = \min\{n : Z_{n-d+1} + \cdots + Z_n = 2\}$; that is, the first time that two ones

appear among the last d observations. For nested sampling plans it is possible to evaluate exact expressions for the ARL and $E_{F(x|\theta_2)}T$. We demonstrate such calculations for the multi-normal case, and show that for this example the speed of detection of a change is close to that of the CUSUM procedure.

Theorem 53.2.1 *For* $\Pi^2(d;2)$ *the expectation and standard deviation of the stopping time* T *are given by*

$$E(T \mid \Pi^2(d;2); \quad \theta_h) \equiv \mathbf{E_h}(\mathbf{T}) = \frac{n\left(2 - {P_h}^{d-1}\right)}{Q_h\left(1 - {P_h}^{d-1}\right)}, \tag{53.2.4}$$

$$\sigma(T \mid \Pi^2(d;2); \quad \theta_h) \equiv \sigma_{\mathbf{h}}(\mathbf{T}) = \frac{n\left[2P_h + {P_h}^{2d-1} + {P_h}^{d-1}((2d+1)Q_h - 2)\right]^{0.5}}{Q_h\left(1 - {P_h}^{d-1}\right)}. \tag{53.2.5}$$

53.3 Multivariate Normal and One-parameter Exponential Distributions

We consider a situation where the observations $(X_{i1}, X_{i2}, \ldots X_{in})$, $i = 1, 2, \ldots$ have the k-variate normal distributions with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2. If $\Sigma_1 = \Sigma_2 \equiv \Sigma$ then the logarithm of the likelihood ratio (from 53.1.1) is given by

$$Y_i(\mu_1, \mu_2, \Sigma) = n(\mu_2 - \mu_1)'\Sigma^{-1}\bar{X}_i - 0.5n(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1), \tag{53.3.6}$$

where $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$.

Theorem 53.3.1 *Let* μ_1, μ_2 *and* Σ *be known. Then for any* μ_h *the probability (53.1.3) is given by the following formula:*

$$P_h = P(Z_i = 0 \mid \mu_h) = \Phi\left(\frac{C + 0.5n(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1 - 2\mu_h)}{\sqrt{nR^2}}\right). \tag{53.3.7}$$

Here $R^2 = (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)$ *and* $\Phi(x)$ *is the the standard normal distribution function.*

Corollary 53.3.1

$$P_1 = \Phi\left(\frac{C}{\sqrt{nR^2}} + 0.5\sqrt{nR^2}\right), \quad P_2 = \Phi\left(\frac{C}{\sqrt{nR^2}} - 0.5\sqrt{nR^2}\right). \tag{53.3.8}$$

Note that the probabilities P_1 and P_2 do not depend on the parameters μ_1, μ_2, and Σ. These probabilities depend only on R^2. If $\mu_h \neq \mu_1$, $\mu_h \neq \mu_2$ then according to (53.3.7) P_h will depend on all the parameters μ_h, μ_1, μ_2, Σ.

We consider also the one-parameter exponential families of distributions with density function

$$f(x \mid \theta) = u(x) \exp\{a(\theta)s(x) + b(\theta)\}. \tag{53.3.9}$$

Here $a(\theta)$ and $b(\theta)$ are continuous functions of the parameter θ, $\theta \in \mathbf{\Theta} \subset \mathbf{R^1}$, $x \in \mathbf{A} \subset \mathbf{R^1}$. Normal, Rayleigh, Pareto, Weibull and other one-parameter families of distributions have such density functions (53.3.9).

According to the formula (53.1.1) in this case

$$Y_i(\theta_1, \theta_2) = \ln \prod_{j=1}^{n} \frac{f(X_{ij} \mid \theta_2)}{f(X_{ij} \mid \theta_1)} = S_i(n)[a(\theta_2) - a(\theta_1)] + n[b(\theta_2) - b(\theta_1)], \tag{53.3.10}$$

where $S_i(n) = \sum_{j=1}^{n} s(X_{ij})$ is the sufficient statistic for the family of distributions (53.3.9). Using the distribution of the sufficient statistic $S_i(n)$ the probability (53.1.3) has the form

$$P_h = P(Y_i(\theta_1, \theta_2) \leq C \mid \theta_h) = P\left(S_i(n)[a(\theta_2) - a(\theta_1)] + n[b(\theta_2) - b(\theta_1)] \leq C \mid \theta_h\right). \tag{53.3.11}$$

Example 53.3.1 *Let the random variable X_{ij} have the one-parameter Pareto distribution (λ is known) with density function*

$$f(x|\theta) = \frac{\theta\ \lambda^{\theta}}{x^{\theta+1}} = \frac{1}{x} \exp\{-\theta \ln x + \ln(\theta\lambda^{\theta})\}; \quad x \geq \lambda > 0; \quad \theta > 0. \tag{53.3.12}$$

In this case the sufficient statistic is $S_i(n) = \sum_{j=1}^{n} \ln(X_{ij})$. If $\theta_1 > \theta_2$ then the probability (53.3.11) can be written in the form

$$P_h = P(Y_i(\theta_1, \theta_2) \leq C \mid \theta_h) = G\left(\theta_h \left([C + n \ln(\theta_2 {\theta_1}^{-1} \lambda^{\theta_1 - \theta_2}](\theta_1 - \theta_2)^{-1} - n \ln \lambda\right)\right). \tag{53.3.13}$$

Here $G(x) = \int_o^x t^{n-1} e^{-t} dt$ is the cumulative distribution function of the Gamma distribution.

53.4 Numerical examples and comparisons

Example 53.4.1 *Consider the nested plan with $n = 4$, $d = 3$ and the X_{ij} have tri-variate normal distributions with in-control and out-of-control means μ_1, μ_2 respectively, and covariance matrix Σ given by:*

$$\mu_1 = (0, 0.5, 0.7)'\,;\ \mu_2 = (0.8, 1.7, 1.5)'\,;\ \Sigma = diag(2, 4, 2).$$

Table 53.1: Out-of-control mean and standard deviation of the run length of nested plans when the initial and the final distributions are multivariate normal with known different means and the same covariance matrix

$\mathbf{E_1(T)}$	**d**	**R**	**n**	$\mathbf{E_2(T)}$	$\mathbf{P_1}$	**C**	$\mathbf{P_2}$	$\sigma_1\mathbf{(T)}$	$\sigma_2\mathbf{(T)}$
1000	3	0.5	8	40.58	0.9315	1.1032	0.5291	985.42	28.53
1000	3	0.5	16	43.12	0.8999	0.5618	0.2360	971.79	18.01
1000	4	0.5	6	43.56	0.9511	1.2773	0.6666	986.56	33.22
1000	4	0.5	12	40.86	0.9281	1.0324	0.3936	974.17	21.07
1000	3	2.0	1	4.64	0.9770	1.9909	0.4982	998.07	3.14
1000	3	2.0	4	8.08	0.9527	-1.3137	0.0099	992.52	0.58
1000	4	2.0	1	4.63	0.9811	2.1524	0.5304	997.61	2.97
1000	4	2.0	4	8.10	0.9607	-0.9644	0.0125	990.87	0.64
2000	3	0.5	8	51.54	0.9527	1.3640	0.6016	1985.05	39.39
2000	3	0.5	16	48.32	0.9315	0.9744	0.3040	1970.84	23.71
2000	4	0.5	6	57.72	0.9663	1.4901	0.7272	1986.15	47.09
2000	4	0.5	12	47.84	0.9511	1.3670	0.4694	1973.11	27.98
2000	3	2.0	1	5.52	0.9839	2.2825	0.5562	1998.05	4.01
2000	3	2.0	4	8.13	0.9671	-0.6410	0.0154	1992.38	0.73
2000	4	2.0	1	5.45	0.9868	2.4377	0.5866	1997.58	3.77
2000	4	2.0	4	8.15	0.9728	-0.3051	0.0189	1990.63	0.79

Table 53.2: Out-of-control ARL of the multivariate CUSUM parametric procedures (by simulation) and those of the proposed nested plan for detecting a change in the mean of the bivariate normal distribution

R	$\mathbf{E_1(T)}$	**Rule**	$\mathbf{E_2(T)}$
1	200	$CUSUM$	9.35
1	200	$\prod_{nes}(n=2;d=3)$	9.43
2	200	$CUSUM$	3.48
2	200	$\prod_{nes}(n=1;d=3)$	3.29
3	200	$CUSUM$	1.69
3	200	$\prod_{nes}(n=1;d=3)$	2.19

We thus obtain

$$R^2 = (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) = \frac{0.8^2}{2} + \frac{1.2^2}{4} + \frac{0.8^2}{2} = 1$$

For example for setting the ARL at $E_1(T) \equiv E_{F(x|mu_1,\ \Sigma)}(T) = 1000$, *we use (53.2.4) and* $n = 4$, $d = 3$ *to obtain*

$$\frac{4\left(2 - {P_1}^2\right)}{Q_1 P_1\left(1 - {P_1}^2\right)} = 1000,$$

and hence $P_1 = 0.95270$. *By (53.3.8) we get* $C = 1.343152$ *and* $P_2 = 0.37130$. *According to the formulas (53.2.4) and (53.2.5)* $E_2(T) = 13.74$, $\sigma_1(T) = 992.52$, $\sigma_2(T) = 7.67$. *If* $\mu_3 = (1.4\ 2.6\ 2.1)'$ *then from (53.3.7) the probability* P_3 *is*

$$P_3 = \Phi\left(\frac{1.343152 + 2(0.8\ \ 1.2\ \ 0.8)'\Sigma^{-1}(-2\ \ -3\ \ -2)}{2}\right) \simeq \Phi(-1.83) = 0.034.$$

In Table 53.1 we provide the results of computing, using (53.2.4, 53.2.5) and (53.3.8), some values for the out-of-control mean and the standard deviation of the run length for different (given) values of in-control ARL and for various values of d, n, *and* R. *From Table 53.1 we conclude that for a given value of*

in-control ARL the influence of d on the out-of-control ARL is smaller than the influence of n. Obviously, increasing R yields a decrease in the out-of-control ARL. In addition, it is possible to see that for small values of R large values of n are required in order to decrease the out-of-control ARL. We conclude also that the standard deviation of the in-control run length is a little smaller than the corresponding mean run length but, roughly, the values are quite similar. Note that the optimal choice of d and n for the proposed nested plan, in order to minimize the out-of-control ARL, remains an important open question and demands further special consideration.

The change point problem for multivariate normal observations was also considered by Crosier (1988) in the context of the CUSUM procedure. We compare in Table 53.2 the nested plan and the multivariate CUSUM procedure (see Crosier (1988))for detecting a change in the mean of the bivariate normal distribution, assuming all parameters (μ_1, μ_2 and Σ) are known.

References

1. Crosier, R. B. (1988), Multivariate Generalizations of Cumulative Sum Quality-Control Schemes, *Technometrics,* **30**, 291-303.

2. Feigin, P., Gurevich, G., Lumelskii, Ya. (2005). Nested Plans and the Sequential Change Point Problems, *Proceedings of the International Symposium on Stochastic Models in Reliability, Safety Security and Logistics,* Beer Sheva, Israel, Abstracts book, 107-110.

3. Lai, T. L. (1995), Sequential changepoint detection in quality control and dynamical systems, *Journal R. Statist. Soc. B,* **57**, 1-33.

4. Lumelskii, Ya. P., Feigin, P. D. (2005) Embedded polynomial plans of random walks and their applications, *Journal of Mathematical Sciences*, **127**, 2103-2113.

5. Lumelskii, Ya., Gurevich, G., Feigin, P. D.(2006). Nested Plans as Sequential Quality Control Schemes for Detecting a Change in a Multivariate Normal Distribution, *Quality Technology and Quantitative Management,* accepted for publication.

6. Zacks, S. (1991), Detection and change-point Problems, Handbook of Sequential Analysis, Statist. Textbooks Monogr., Dekker, New York, V. 118, 531-562.

54

Incorporating Bayesian Models For The Estimation Of The Spread Parameters Of Probabilistic Neural Networks With Application In Biomedical Tasks

V.L. Georgiou[1], **S. Malefaki**[2]

Department of Mathematics, University of Patras, Greece[1]
Department of Statistics and Insurance Science, University of Piraeus, Greece[2]

Abstract:

One of the tools that have been implemented for classification tasks in biomedical applications is the Probabilistic Neural Network (PNN). In order for the PNN to work adequately, an optimum selection of spread parameters of the PNN must be made. In this contribution, two bayesian models are proposed for the selection of the spread parameters. The mean of the posterior distribution of the spread parameters is used in the PNN. The proposed approach is applied to three biomedical real–world datasets and the obtained results are compared with the ones obtained from feed–forward neural networks.

Keywords and phrases: Probabilistic neural networks, Spread parameter

54.1 Introduction

During the past few years there has become a rapid development of the research in bioinformatics and medical tasks. Some of the tools that have been used extensively in these fields of science are Neural Networks, and especially Probabilistic Neural Networks (PNN). PNNs have been implemented to perform cancer classification in Huang (2002). Also, PNNs are employed to develop accurate NMR-based metabonomic models for the prediction of xenobiotic-induced toxicity in experimental animals and their possible future use in accelerated drug discovery programs is highlighted in Holmes *et al.* (2001).

PNNs were introduced by Specht (1990) and constitute a class of neural networks that combine some of the best attributes of statistical pattern recog-

nition and feed–forward neural networks. PNNs are the neural network implementation of kernel discriminant analysis. In contrast to feed–forward neural networks that are black-box systems, PNNs use the Bayes decision rule for pattern classification. A PNN not only classifies a new pattern but also provides a measure of the uncertainty of that classification, since it provides the class posterior probabilities given the new pattern.

One drawback of the performance of PNNs is the need to estimate a promising spread parameter of the network's kernel. This is usually obtained by a trial-and-error procedure, although several alternative methods have been proposed in the literature. Georgiou *et al.* (2006), Gorunescu *et al.* (2005). In this contribution, a new approach for the estimation of the PNN'S spread parameters is proposed. We confront this task by a Bayesian approach. We model the centered data using simple bayesian models with conjugated priors and take for spread parameter the mean of its posterior distribution. The approach has been implemented on three biomedical datasets from the UCI repository with encouraging results.

54.2 Probabilistic Neural Networks

The PNN was introduced by Specht as a new neural network type, although it was already widely known in the statistical literature as kernel discriminant analysis, Hand (1982). What Specht introduced was the neural network approach of kernel discriminant analysis which incorporates the Bayes decision rule and the non–parametric density function estimation of a population, Parzen (1962).

The training procedure of the PNN requires only a single pass of the patterns of the training data, which results in a very small training time. In fact, the training procedure is just the construction of the PNN from the available data. The structure of the PNN has always four layers; the *input layer*, the *pattern layer*, the *summation layer*, and the *output layer*. An input feature vector, $X \in \mathbb{R}^p$, is applied to the p input neurons and is passed to the pattern layer. The pattern layer is organized into K groups, where K is the number of classes present in the data set. Each group of neurons in the pattern layer consists of N_k neurons, where N_k is the number of training vectors that belong to class k, $k = 1, \ldots, K$. The ith neuron in the kth group of the pattern layer computes its output using a kernel of the form,

$$f_{ik}(X) = \exp\left(-\frac{1}{2}\left(X - X_{ik}\right)^T \Sigma_k^{-1}\left(X - X_{ik}\right)\right), \tag{54.2.1}$$

where $X_{ik} \in \mathbb{R}^p$ is the center of the kernel and Σ_k is the matrix of spread (smoothing) parameters of the kernel. The summation layer has K neurons

and estimates the conditional class probabilities as

$$G_k(X) \propto \sum_{i=1}^{N_k} \pi_k f_{ik}(X), \quad k \in \{1, \ldots, K\}, \tag{54.2.2}$$

where π_k is the prior probability of class k, $\sum_{k=1}^{K} \pi_k = 1$. So a vector X is classified to the class that has the maximum output of its summation neuron. Another variant of the PNN is to train it not by using the whole training data set but a part of it. Such a training set can be obtained either by randomly sampling from the available data or by finding some "representatives" of the training data through a clustering technique. In this implementation, we identified some informative representatives (mean centers) from each class by using the K-means clustering algorithm , MacQueen (1967), on the training data of each class. So, instead of using all the available training data, we extracted a few centers from each class and used these vectors as centers for the kernels of the PNN. The number of centers we extracted using K-means was the 10% of the size of each class. This resulted in a PNN with size ten times smaller than the PNN using all the available training data as centers of its kernels.

As we mentioned earlier, a spread parameter must be set to the PNN in order to achieve an adequate performance. It is assumed that each class has its own matrix of spread parameters $\Sigma_k = \text{diag}(\sigma_{1k}^2, \ \ldots, \sigma_{pk}^2)$, $k = 1, \ldots, K$.

In the next section, a new approach to the task of identifying Σ_k is presented.

54.3 Proposed Approach

We consider two different bayesian models for each dimension of the centered data in each one of the K classes, since it is assumed that the matrix of spread parameters is diagonal.

First, we consider the following simple model:

$$\begin{aligned} X_{ik} &\overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, \Sigma_k), \quad i = 1, \ldots, N_k, \\ \tau_{jk} &\overset{\text{iid}}{\sim} \mathcal{G}(\alpha, \beta), \quad j = 1, \ldots, p, \end{aligned}$$

where $\tau_{jk} = \sigma_{jk}^{-2}$ and $\alpha, \beta > 0$ are known parameters.

The posterior distribution of τ_{jk} given the data is:

$$\tau_{jk} | X_{jk} \sim \mathcal{G}\left(\frac{N_k}{2} + \alpha, \frac{\sum_{i=1}^{N_k} X_{ijk}^2}{2} + \beta \right).$$

Hence, τ_{jk} is obtained by the mean of the posterior distribution. So, we use for the spread parameter, the quantity $\frac{\sum_{i=1}^{N_k} X_{ijk}^2/2 + \beta}{N_k/2 + \alpha}$.

Second, we consider the following two–parameter bayesian model:

$$\begin{aligned} X_{ik} &\overset{\text{iid}}{\sim} \mathcal{N}_p(\mu_k, \Sigma_k) \quad i = 1, \ldots, N_k, \\ \mu_{jk} &\sim \mathcal{N}(0, 1), \\ \tau_{jk} &\sim \mathcal{G}(\alpha, \beta), \quad j = 1, \ldots, p, \end{aligned}$$

where α, β are known parameters. It is assumed that X_{ik} are conditionally independent given μ_k, τ_{jk} and μ_k and τ_{jk} are themselves independent. The joint posterior distribution of μ_{jk} and τ_{jk} is:

$$\pi(\mu_{jk}, \tau_{jk}|X_{jk}) \propto \tau^{N_k/2+\alpha-1} \exp\left(-\tau_{jk}\left(\frac{\sum_{i=1}^{N_k}(X_{ijk}-\mu_{jk})^2}{2}+\beta\right)-\frac{\mu_{jk}^2}{2}\right)$$

In order to estimate τ_{jk}, we use Gibbs sampler, Geman and Geman (1984), with transition kernel the product of the full conditional for μ_{jk} and τ_{jk}, which produces a markov chain with stationary distribution the posterior of them. In these models, the prior distributions are conjugated to the likelihood so the full conditional distributions reduce analytically to closed form distributions. It is not of great importance to choose conjugated prior distributions to the likelihood. Any distribution can be chosen for prior. In these cases, we can use other Monte Carlo or Markov Chain Monte Carlo simulation methods such as Importance Sampling, Metropolis Hastings in order to estimate τ_{jk}, Gilks *et al.* (1996).

54.4 Experimental Results

The proposed approach has been applied to three biomedical problems. The three datasets come from the UCI data repository and are implemented according to the Proben1 rules, Prechelt (1994). The results are compared with the ones obtained by feed–forward neural networks from Proben1. The first data set is "The Breast Cancer Data Set". There are two possible classes for each record: benign or malignant. The input features are the uniformity of cell size and shape; bland chromatin; single epithelial cell size; and mitoses. There are 9 continuous inputs and 699 instances. Also, there are no missing values.

The second is "The Pima Indians Diabetes Data Set". It concerns the Pima Indians diabetes and the input features are the diastolic blood pressure; triceps skin fold thickness; plasma glucose concentration in a glucose tolerance test; and diabetes pedigree function. The 8 inputs are all continuous without missing values and there are 768 instances. The aim is to classify whether someone is infected by diabetes or not, therefore, there are two classes.

The aim of the third dataset, named "Heart Disease", is to decide whether at least one of four major vessels of the heart is reduced in diameter by more

Cancer			
α	β	Model 1	Model 2
0.001	0.001	1.15	1.15
0.01	0.01	1.15	1.15
0.1	0.1	1.72	1.15
0.01	0.1	1.72	1.15

Diabetes		
α	β	Model 1
0.001	0.001	26.56
0.01	0.01	26.56
1	1	25.52
3	1	25.52

Heart			
α	β	Model 1	Model 2
0.01	0.01	19.13	20.44
0.1	0.1	18.69	20.44
0.001	3.65	21.73	17.83

Proben1 results		
Dataset	Mean	St.Dev
Cancer	1.47	0.60
Diabetes	24.57	3.53
Heart	19.89	2.27

Table 54.1: Misclassification Proportions (%) using PNN for several parameters α and β and Proben1 Neural Networks

than 50%. The binary decision is made based on personal data such as age, sex, smoking habits, subjective patient pain descriptions and results of various medical examinations such as blood pressure and electro cardiogram results. There are 35 inputs and 920 instances. The misclassification proportions of the proposed approach of the PNNs for the three medical datasets are presented in Table 54.1. They are presented for each one of the two proposed models for several values of α and β and compared with the ones obtained by neural networks. In two out of the three datasets, PNNs have reached a better performance compared to neural networks performance obtained by Proben1.

It must also be noted that the performance of PNNs implemented by the proposed approach is quite robust to the selection of parameters α and β and especially to the parameter α. Another important advantage of the proposed approach of PNNs is the robustness of the estimation of the spread parameters given α and β. The estimation is quite robust to the initial values compared to the initialization of feed–forward neural networks that don't always converge to the same values of parameters. That is why only one value of the misclassification proportion of the PNNs is presented compared to neural networks where the mean and standard deviation of the misclassification proportions is presented.

54.5 Conclusion

PNNs have been widely used in various fields of science. In order for the PNN to achieve an adequate performance, a good choice of spread parameters must be made. In this contribution a new approach for the estimation of the spread parameters of a PNN is proposed. This approach incorporates two bayesian models for the estimation of the spread parameters. This approach is applied

to three real-world biomedical datasets, namely breast cancer, diabetes and heart disease with encouraging results. The results are compared with the ones obtained by feed–forward neural networks and in breast cancer and heart disease datasets, the performance of PNNs is superior.

Acknowledgment

We thank European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II) and particularly the Program IRAKLEITOS for funding the above work.

References

1. Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Trans. Pattn. Anal. Mach. Intel.*, **6**, 721–741.
2. Gorunescu, M., Gorunescu, F., Ene, M. and El-Darzi, E. (2005). A heuristic approach in hepatic cancer diagnosis using a probabilistic neural network-based model, In *Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis*, pp 1016–1024, Brest, France.
3. Georgiou, V. L., Pavlidis, N. G., Parsopoulos, K. E., Alevizos, Ph. D. and Vrahatis, M. N. (2006). New self-adaptive probabilistic neural networks in bioinformatic and medical tasks, *Int. Journal on Artificial Intelligence Tools, to appear.*
4. Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall.
5. Hand, J. D. (1982). *Kernel Discriminant Analysis*, Research Studies Press, Chichester.
6. Holmes, E., Nicholson, J. K. and Tranter, G. (2001). Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks, *Chem. Res. Toxicol.*, **14**(2), 182–191.
7. Huang, C. J. (2002). A performance analysis of cancer classification using feature extraction and probabilistic neural networks, In *Proceedings of the 7th Conference on Artificial Intelligence and Applications*, pp 374–378.
8. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations, In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp281–297, Berkeley.
9. Parzen, E. (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **3**, 1065–1076.
10. Prechelt, L. (1994). Proben1: A set of neural network benchmark problems and benchmarking rules, Technical Report 21/94, Fakultät für Informatik, Universität Karlsruhe.
11. Specht, D. F. (1990). Probabilistic neural networks, *Neural Networks*, **1**(3), 109–118.

55

Estimation with Unknown Selection Bias and Censoring

Agathe Guilloux

LSTA Université Paris 6, 175 rue du Chevaleret, 75013 PARIS, aguillou@ccr.jussieu.fr

Abstract: We consider the problem of estimating the distribution of a pair of random variables (X, σ), where X represents a lifetime and σ a birth time, when the pair (X, σ) is observed only conditionally on being in a Borel subset $\mathcal{S}$ of $\mathbb{R} \times \mathbb{R}_+$. Most of the sampling pattern in survival analysis can be described via a set $\mathcal{S}$. Considering in addition the possibility of random censoring, we construct estimators for the distribution functions of X and σ based on our biased sample. We show weak convergence theorems for both estimators.

Keywords and phrases: Selection bias, Product-limit estimation, Martingale, Gaussian process, Weak convergence

55.1 Introduction

Consider a population of individuals $i \in I$. Let the random variable (r.v.) σ_i be the birth date of individual i and the non-negative r.v. X_i its lifetime. As described by Keiding (1990) and Lund (2000), consider a coordinate system with the calendar time as abscissa and the age as ordinate, it will be referred to as the Lexis diagram (Lexis, 1875). In this diagram, a life-line $\mathcal{L}(\sigma, X)$ is defined by:

$$\mathcal{L}(\sigma, X) = \{(\sigma + y, y), 0 \leq y \leq X\}.$$

The population I is then represented in the Lexis diagram as the set of all life-lines $\mathcal{L}(\sigma_i, X_i)$ for $i \in I$.

In classical survival analysis, one would consider that an i.i.d. sample (possibly censored) from population I could be drawn and then would estimate the distribution of X on the basis of this i.i.d. sample. In practice however, it may happen that the way individuals are chosen for the study prevent from observing a i.i.d. sample directly from population I.

As carefully described by Lund (2000), most of the sampling patterns in survival analysis can be described as follows. Let $\mathcal{S}$ be a deterministic Borel set in the Lexis diagram. Consider now that only the individuals whose life-lines intersect the Borel set $\mathcal{S}$ can be included in the study, i.e. the individuals with a pair (σ, X) such that $\mathcal{L}(\sigma, X) \cap \mathcal{S} \neq \emptyset$.

Let $\sigma_{\mathcal{S}}$ denote the birth time and $X_{\mathcal{S}}$ the lifetime for the included individuals. For now on, the pair $(\sigma_{\mathcal{S}}, X_{\mathcal{S}})$ will be referred to as the observable r.v. as opposed the unobservable pair (σ, X).We show in Section 2 that, under some condition on the collection $(\sigma_i)_{i\in I}$, we have, for all $t \geq 0$:

$$F_{\mathcal{S}}(t) = \mathbb{P}\left(X_{\mathcal{S}} \leq t\right) = \frac{\int_{[0,t]} w(v)dF(v)}{\mu_{\mathcal{S}}}, \tag{55.1.1}$$

where F is the distribution function (d.f.) of the r.v. X and w is a non-negative weight function, which depends only on the distribution of the r.v. σ and $\mu_{\mathcal{S}} = \int_0^\infty w(v)dF(v)$.

The problem addressed here is to estimate the d.f. F of the r.v. X and the weight function w on the basis of an i.i.d. censored (in a way to be defined later) sample of $(\sigma_{\mathcal{S}}, X_{\mathcal{S}})$.

The r.v. $X_{\mathcal{S}}$ with d.f. $F_{\mathcal{S}}$ given in Equation (55.1.1) is usually said to suffer from a selection bias. In the case where the weight function w is known, the problem of estimating the cumulative distribution function (c.d.f.) F of X given an i.i.d. biased sample $X_{\mathcal{S},1}, \ldots, X_{\mathcal{S},n}$ has received a lot of attention. We refer to Gill *et al.*(1988) and Efromovich (2004) for theoretical results in the general case. The special case where $w(x) = x$ for all $x > 0$, called "length-biased sampling", has received a particular attention, see Vardi (1982), de Uña-Àlvarez (2002,2004) and Asgharian *et al.* (2002). Unfortunately these results cannot be applied here as w is not assumed to be known.

On the other hand, Winter and Fldes (1988) have constructed and studied a product-limit type estimator of the d.f. F on the basis of a censored biased sample of $(\sigma_{\mathcal{S}}, X_{\mathcal{S}})$, without assuming that w is known. They still considered the particular case where $\mathcal{S} = \{(t_0, y), y \geq 0\}$ and a deterministic censoring.

55.2 Sampling in the Lexis diagram

55.2.1 Modeling the Lexis diagram

Consider the Lexis diagram for a population of individuals $i \in I$ as described in Section 1 and a Borel set $\mathcal{S}$ in $\mathcal{B}_{\mathbb{R}\times\mathbb{R}_+}$ (the Borel σ-algebra on $\mathbb{R} \times \mathbb{R}_+$) describing the sampling pattern. As mentioned earlier, an individual i in the population, with birth date σ_i and lifetime X_i, is included in the sample if its life-line $\mathcal{L}(\sigma_i, X_i)$ intersects the Borel set $\mathcal{S}$.

Let the age $a_{\mathcal{S}}(s)$ at inclusion for the birth time s in $\mathbb{R}$ be defined as:

$$\begin{cases} a_{\mathcal{S}}(s) = \inf\{y \geq 0, (s+y, y) \in \mathcal{S}\} \\ a_{\mathcal{S}}(s) = \infty \text{ if the infinimum does not exist.} \end{cases}$$

The individual i with birth date σ_i and lifetime X_i is then included in the sample if:

$$\mathcal{L}(\sigma_i, X_i) \cap \mathcal{S} \neq \emptyset \Leftrightarrow a_{\mathcal{S}}(\sigma_i) < \infty \text{ and } X_i \geq a_{\mathcal{S}}(\sigma_i). \tag{55.2.2}$$

Now, following Lund (2000), we assume that the point process $\eta = \sum_{i \in I} \varepsilon_{\sigma_i}$, with the collection of birth times as occurrence times, is a non-homogeneous Poisson process on $\mathbb{R}$ with intensity φ (where ε_a is the Dirac measure at point a). We assume furthermore, that the lifetimes X_i, for $i \in I$, are i.i.d. with common probability density (p.d.f.) function f.

The properties of Poisson processes assure that we have, for all $s \in \mathbb{R}$ and $t \in \mathbb{R}_+$:

$$\begin{aligned} & \mathbb{P}\left(\sigma_{\mathcal{S}} \leq s, X_{\mathcal{S}} \leq t\right) \qquad (55.2.3) \\ = & \frac{\int\int_{]-\infty,s]\times[0,t]} I\left(\{a_{\mathcal{S}}(u) < \infty\}\right) I\left(\{a_{\mathcal{S}}(u) \leq v\}\right) \varphi(u) f(v) du dv}{\mu_{\mathcal{S}}}, \end{aligned}$$

where

$$\mu_{\mathcal{S}} = \int\int_{\mathbb{R}\times\mathbb{R}_+} I\left(\{a_{\mathcal{S}}(u) < \infty\}\right) I\left(\{a_{\mathcal{S}}(u) \leq v\}\right) \varphi(u) f(v) du dv.$$

Hence the marginal distribution of the r.v. $X_{\mathcal{S}}$ is given, for all $t \in \mathbb{R}_+$, by:

$$F_{\mathcal{S}}(t) = \mathbb{P}\left(X_{\mathcal{S}} \leq t\right) = \frac{1}{\mu_{\mathcal{S}}} \int_0^t w(v) f(s) ds, \tag{55.2.4}$$

with

$$w(t) = \int_{-\infty}^{\infty} I\left(\{a_{\mathcal{S}}(u) \leq t\}\right) \varphi(u) du. \tag{55.2.5}$$

On the other hand, the marginal distribution of the r.v. $\sigma_{\mathcal{S}}$ is given, for all $s \in \mathbb{R}$, by:

$$\Phi_{\mathcal{S}}(s) = \mathbb{P}(\sigma_{\mathcal{S}} \leq s) = \frac{1}{\mu_{\mathcal{S}}} \int_{-\infty}^{s} \varphi(u) \bar{F}(a_{\mathcal{S}}(u)) du, \tag{55.2.6}$$

where $\bar{F} = 1 - F$. Our aim is then to estimate the functions F and w on the basis of a biased (censored) sample from $(\sigma_{\mathcal{S}}, X_{\mathcal{S}})$.

55.2.2 Censored observations

Now only the individuals, whose life-lines intersect the Borel set $\mathcal{S}$, are included in the study. For included individual i, with birth date $\sigma_{\mathcal{S},i}$ and lifetime $X_{\mathcal{S},i}$, we assume that its age at inclusion $a_{\mathcal{S}}(\sigma_{\mathcal{S},i})$ is observable. The lifetime $X_{\mathcal{S},i}$ can straightforwardly be written as follows:

$$X_{\mathcal{S},i} = \underbrace{a_{\mathcal{S}}(\sigma_{\mathcal{S},i})}_{\text{age at inclusion}} + \underbrace{(X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}))}_{\text{time spent in the study}}.$$

As the time spent in the study is given by $X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i})$, we shall assume that this time can be censored. It would indeed be the case, for example, if individual i leaves the study before its death. We follow here Asgharian (2003) and Winter and Fldes (1988).

For that matter, we introduce a non-negative r.v. C with d.f. H and independent of $X_{\mathcal{S}}$ and $a_{\mathcal{S}}(\sigma_{\mathcal{S}})$, such that the observable time for individual i is $Z_i = a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) + (X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i})) \wedge C_i$. As usual, we assume furthermore that the r.v. $I\left(X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \leq C\}\right)$ (where $I(.)$ is the indicator function) is observable. As a consequence, the available data consists for $i = 1, \ldots, n$ in:

$$\begin{cases} \sigma_{\mathcal{S},i} \\ Z_i = a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) + (X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i})) \wedge C_i \\ I\left(\{X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \leq C_i\}\right) \end{cases}.$$

We seek to estimate the d.f. F of the unbiased r.v. X as well as the weight function w defined in Equation (55.2.5) with the data described above.

55.3 Inference for the distribution of the r.v. X

Considering the situation of interest described in Section 55.2, we now introduce the counting process D defined, for all $t \geq 0$, as follows:

$$D(t) = \sum_{i=1}^{n} I\left(\{Z_i \leq t, X_{\mathcal{S},i} - a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \leq C_i\}\right). \tag{55.3.7}$$

Notice that, for $t \geq 0$, the r.v. $D(t)$ is the "number of observed deaths before age t" in the sample. Let furthermore the process O be defined, for all $t \geq 0$, by:

$$O(t) = \sum_{i=1}^{n} I\left(\{a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \leq t \leq X_{\mathcal{S},i}, t \leq a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) + C_i\}\right) \tag{55.3.8}$$

The r.v. $O(t)$ represents the "number of individuals at risk at age t".

Mimicking the construction of the Kaplan-Meier estimator in classical survival analysis, we define the estimator $\widehat{F}_n$ for the d.f. F of the r.v. $X.$, for all $t \geq 0$, by:

$$\hat{F}_n(t) = 1 - \prod_{s \leq t} \left(1 - \frac{dD(s)}{O(s) + n\epsilon_n}\right) \tag{55.3.9}$$

where $\prod$ is the product-integral (see Andersen *et al.* - 1993 - for a review on this topic) and $(\epsilon_n)_{n \geq 1}$ is a sequence of positive numbers such that $\epsilon_n \to 0$ as $n \to \infty$.

Theorem 55.3.1 *Let τ be defined as $\tau = sup\{t > 0, (1 - F(t))(1 - H(t)) > 0\}$. The following convergence holds, for all $\tau' \leq \tau$, as n goes to infinity:* $\sup_{t \leq \tau'} |\hat{F}_n(t) - F(t)| \xrightarrow{\mathbb{P}} 0$.
Moreover the following weak convergence holds in the space $\mathbb{D}[0, \tau']$ of càdlàg functions on $[0, \tau']$ as n goes to infinity: $\sqrt{n}(\hat{F}_n - F) \xrightarrow{\mathcal{D}} L$, where L is a gaussian process with mean zero and variance function given, for all $(s, t) \in [0, \tau']^2$, by:

$$\frac{< L(s), L(t) >}{(1 - F(s))(1 - F(t))} = \int_0^{s \wedge t} \frac{dF(x)}{(1 - F(x))^2 \theta(H, w)(x)},$$

where, for $t \geq 0$: $\theta(H, w)(t) = (1/\mu_w)w(t) - \int_0^t w(t - c)dH(c)$.

55.4 Inference for the weight function

The weighting function w has been defined, for all $t \geq 0$, by:

$$w(t) = \int_{-\infty}^{\infty} I\left(\{a_{\mathcal{S}}(u) \leq t\}\right) \varphi(u) du$$

and Equation (55.2.6) states that $\Phi_{\mathcal{S}}(s) = \mathbb{P}(\sigma_{\mathcal{S}} \leq t) = (1/\mu_{\mathcal{S}}) \int_{-\infty}^{t} \varphi(u)(1 - F)(a_{\mathcal{S}}(u))du$. Hence, for all $t \geq 0$, we have:

$$\frac{w(t)}{\mu_{\mathcal{S}}} = \int_{-\infty}^{\infty} \frac{I\left\{a_{\mathcal{S}}(u) \leq t\}\right)}{(1 - F)(a_{\mathcal{S}}(u))} d\Phi_{\mathcal{S}}(u).$$

An natural estimator for the function $w/\mu_{\mathcal{S}}$ based on the i.i.d. sample described in Section 55.2.2 is then given by:

$$\widehat{\frac{w(t)}{\mu_{\mathcal{S}}}} = \frac{1}{n} \sum_{i=1}^{n} \frac{I\left\{a_{\mathcal{S}}(\sigma_{\mathcal{S},i}) \leq t\}\right)}{1 - \widehat{F}_n(a_{\mathcal{S}}(\sigma_{\mathcal{S},i}))}, \tag{55.4.10}$$

where $\widehat{F}_n$ has been defined in Equation 55.3.9.

Theorem 55.4.1 *As n goes to infinity, the following weak convergence holds in the space of càdlàg functions $\mathbb{D}[0,\infty]$:*

$$\sqrt{n}\left(\frac{\widehat{w(\cdot)}}{\mu_{\mathcal{S}}}-\frac{w(\cdot)}{\mu_{\mathcal{S}}}\right)\xrightarrow{\mathcal{D}}\int_0^{\cdot}\frac{1}{1-F(a_{\mathcal{S}}(s)}dK(s))+\int_0^{\cdot}\frac{L(s)d\mathbb{P}\left(a_{\mathcal{S}}(\sigma_{\mathcal{S}}\leq s)\right)}{(1-F(a_{\mathcal{S}}(s)))2},$$

where $K(\cdot)=B\circ\mathbb{P}\left(a_{\mathcal{S}}(\sigma_{\mathcal{S}}\leq\cdot)\right)$ is a brownian bridge.

References

1. Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer-Verlag.
2. Asgharian, M., M'Lan, C.E. and Wolfson, D.B. (2002). Length-biased sampling with right censoring: an unconditional approach. *J. Amer. Statist. Assoc.* **97**, 201-209.
3. Asgharian, M. (2003). Biased sampling with right censoring: a note on Sun, Cui & Tiwari (2002). *Canadian Journal of Statistics* **30**, 475-490.
4. Efromovich, S. (2004). Distribution estimation for biased data. *J. Statist. Plann. Inference* **124**, 1-43.
5. Gill, R.D., Vardi, Y. and Wellner, J.A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1172.
6. Keiding, N. (1990). Statistical inference for the Lexis Diagram. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **332**, 487-509.
7. Kingman, J.F.C. (1993). *Poisson processes.* Oxford Science Publications
8. Lexis, W. (1875). Einleitung in die Theorie der Bevölkerung-Statistik. In *Mathematical demography* (édition D. Smith and N. Keyfitz), *Biomathematics* **6**, 39-41 (1977). Springer-Verlag, Berlin.
9. Lund, J. (2000). Sampling bias in population studies - How to use the Lexis diagram. *Scand. J. Statist.* **27**, 589-604.
10. de Uña-Àlvarez, J. (2002). Product-limit estimation for length-biased censored data. *Test* **11**, 109-125.
11. de Uña-Àlvarez, J. (2004). Nonparametric estimation under length-biased sampling and type I censoring: a moment based approach. *Ann. Inst. Statist. Math.* **56**, 667–681.
12. Vardi, Y. (1982). Nonparametric estimation in presence of length bias. *Ann. Statist.* **10**, 616-620.
13. Winter B.B. et Fldes A. (1988) A product-limit estimator for use with length-biased data, Canad. J. Statist., **16** 337-355.

56

Bio-medical Immuno-regulation by Antivirals and their Use in Chronic Infections

Günther Haase

Praxis für Immunotherapie
Germany

Abstract The combination of some specific amino-and nucleic-acids+ enzymes+radical scavenger chinons show a great immuno-stimulation up to immuno-regulation! By this way most of chronic bacterial and viral infections in different organs and tissues and their connected healthproblems can be solved! The patients come back to a normal life, to health and wealth!

A demand of the WHO!

I present also comparative statistics in graphicform!

The GERUSCITH (German-Russian Cooperation for Immunotherapy, Hyperthermia and Neuro-Immuno-Modulation) has proved clinically and scietifically the successful application of BIO-MEDICINE in ultralow dosage of immunoregulative amino-and nucleic-acids.

57

Genetic Epidemiology of Breast Cancer; the Experience in Cyprus

A. Hadjisavvas[1], **M. Loizidou**[1], **A. Adamou**[2], **Y. Markou**[3], **Ch. G. Christodoulou**[4], **K. Kyriacou**[1]

[1]*Department of Electron Microscopy and Molecular Pathology, The Cyprus Institute of Neurology and Genetics*
[2]*Prevention Centre, Nicosia, Cyprus,*
[3]*Bank of Cyprus Oncology Center, Nicosia, Cyprus*
[4]*Department of Molecular Virology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus*

Abstract: Breast cancer is a complex disease and it shows extensive heterogeneity with respect to clinical, histological biological and genetic features. In the mid 1990s two genes, were discovered namely BRCA1 and BRCA2, that predispose to familial cases of the disease. Both these genes are inherited in an autosomal dominant manner and show high penetrance. Germline mutations in these genes substantially increase the risk of breast cancer development, characteristically at an early age. In Cyprus breast cancer is the most frequent malignancy in the female population but no data exist, as yet on the role of the BRCA genes in familial cases. This paper presents the first molecular study on the mutations identified in the BRCA1 and BRCA2 genes in Cypriot families with breast cancer. The entire coding regions of the two breast cancer susceptibility genes BRCA1 and BRCA2 were sequenced in breast cancer patients from 40 Cypriot families with multiple cases of breast and ovarian cancer. In total four protein truncating mutations were found in six families. In BRCA1 a novel truncating mutation 5429delG was found in exon 21. In BRCA2 three truncating mutations were detected; a frameshift 8984delG in exon 22 and two nonsense mutations C1913X in exon 11 and K3326X in exon 27. It is noted that mutation 8984delG was found in 3 unrelated families and haplotype analysis showed that this may be a founder mutation in the Cypriot population.

Keywords and phrases: Familial breast cancer, BRCA mutations, Cypriot families

57.1 Introduction

57.1.1 Breast cancer: incidence, mortality and epidemiology

Breast cancer is the most common cancer in women in the world and in 2000 there were over 1 million new cases; this accounts for 1/5 of the global burden of female cancers world-wide. There are regional differences and geographic in incidence rates and in the more developed countries the age standardized rate is 63 per 100,000 versus 23 per 100,000 in the less developed countries, Parkin (2001). Overall it is estimated that 25% of women with breast cancer will die because of the disease.

The epidemiology of breast cancer has been studied more extensively than any other human disease. A spectrum of risk factors has been identified that increase the risk of developing breast cancer, including duration of estrogen exposure, late first pregnancy and family history. In contrast higher parity and longer duration of lactation lower the risk but other potential risk factors such as smoking, alcohol and fat intake remain controversial, Dumitrescu and Cotarla (2005). However in most women with breast cancer, a specific risk factor cannot be identified.

57.1.2 Breast cancer: familial genetics and genetic epidemiology

The most important risk factor is a family history and it has been recognized for many years that about 15% of breast cancer patients present with a positive family history, Collaborative Group (2001). Hereditary Breast Ovarian Cancer syndrome (HBOC, MIM113705) is the most common form of inherited breast cancer and in the mid 1990stwo breast cancer susceptibility genes were identified. These were named BReast CAncer 1, BRCA1 gene, Miki et al. (1994), and BRCA2 gene, Wooster et al. (1995). Both of these are novel, highly penetrant genes and encode large proteins with pleiotropic cellular functions, Venkitaraman (2002). These two genes are inherited as autosomal dominant and germline mutations in the BRCA1 and BRCA2 genes greatly increase the risk of developing not only breast, and ovarian, but also other types of cancer. Ten years after their discovery a plethora of pathogenic mutations, 800 in BRCA1 and 400 in BRCA2, and numerous other variants have been characterized (see BIC database http://www.nhgri.nih.gov/Intramural_research/Lab_transfer/BIC).

The proportion of HBOC attributable to BRCA1 and BRCA2 mutations is poorly defined and estimates depend upon the population studied, the number of breast and ovarian cancer cases in the family and the mutation detection techniques used. Szabo and King (1997); Neuhausen (1999); Hodgson et al. (2000). In some populations founder mutations have been identified, as in the Ashkenazi-Jews, Struewing et al. (1997) and in the Icelandic population,

Thorlacius et al. (1996). In selected breast cancer families it is estimated that about 20% will have a pathogenic mutation in the BRCA1 or BRCA2 genes, but in founder populations the prevalence may be higher. For ovarian cancer families the figures range between 10-40% depending on the population studied. In addition a significant proportion of early onset breast cancer cases, unselected for family history, will carry a BRCA1 or a BRCA2 mutation and this rises to 20% in women from founder populations, Sanjose et al. (2003). The prevalence of BRCA mutations in the general population is estimated to be between 1 in 500 and 1 in 1000, but again in founder populations, such as the Ashkenazi Jews, this rises to 2.5%. The current average estimates of risk to BRCA1 mutations carriers is 65% for developing breast cancer and 40% for ovarian cancer. The respective risk for BRCA2 mutation carriers is lower, being 45% for breast cancer and 11% for ovarian cancer, Antoniou et al. (2002). In addition a minority of HBOC are due to germline mutations in other genes such as TP53, CHK2, ATM and PTEN, Ingvarsson (2004). Finally a more comprehensive model of inherited breast cancer susceptibility proposes that disease risks are affected by mutations in a small number of genes causing a high risk, as well as larger number of lower risk gene variants interacting together, Antoniou et al. (2002).

57.1.3 Breast cancer in Cyprus

Breast cancer is the most frequent malignancy in Cypriot women with about 300 new cases diagnosed every year. The aim of this work was to determine the contribution of deleterious BRCA1 and BRCA2 mutations in the development of breast/ovarian cancer in Cypriot families with breast and ovarian cancers.

57.2 Materials and Methods

57.2.1 Patients

For this study, 40 Cypriot families with multiple cases of breast / ovarian cancer were selected, from a database containing 1800 consecutive patients, diagnosed with breast cancer. The probands were selected on the basis of having multiple first degree relatives affected with breast/ovarian cancer and all the patients were recruited after signing a consent form. In total, 75 DNA samples, at least one from each family, were collected from affected individuals. In addition DNA samples were obtained from 50 unrelated healthy Cypriots, with no history of breast or ovarian cancer and matched for age and sex to the patients. These controls were used to estimate the population frequency of the detected BRCA1 and BRCA2 variants.

57.2.2 Mutation analysis

Mutation analysis was performed using PCR and sequencing of all exons, as well as intron boundaries of both BRCA genes. The PCR products were sequenced using the same forward and reverse primers, as the ones used for the PCR amplification. Sequencing was carried out using ABI PRISM di-Deoxy Terminator Cycle sequencing kit on an ABI 9700 thermal cycler and an ABI 310 Genetic Analyzer, (Applied Biosystems). Haplotype analysis was carried out on the affected members from 3 families, who were carriers of the 8984 delG mutations in the BRCA2.

57.3 Results

Analysis of BRCA1 and BRCA2 in the 40 Cypriot families, revealed the presence of 18 variants in BRCA1 and 37 variants in BRCA2. The 18 BRCA1 variants include 1 truncating mutation, 8 missense mutations, 3 polymorphisms and 6 intronic variants. The one truncating mutation, 5429delG in exon 21, is novel and was found in a family with 8 breast cancers. The 8 missense mutations are Q356R, P871L, E1038G, S1040N, L1183K, D1344G, S1512I and S1613G. It is noted that missense mutations Q356R and S1512I occur simultaneously in two families.

The 37 BRCA2 variants include 3 truncating mutations, 14 missense mutations, 8 polymorphisms and 12 intronic variants. The 3 truncating mutations include two nonsense mutations, which were found in two separate families. The first is a novel nonsense mutation at codon 1913, in exon 11, a cysteine to a STOP, and the second at codon 3326 in exon 27, a lysine to a STOP. The frameshift mutation 8984delG was detected in five patients from three different families. This mutation appears to be the most frequent deleterious BRCA2 mutation in the families studied so far. Haplotype analysis revealed that this mutation is likely to originate from a common founder in the Cypriot population.

57.4 Discussion

The incidence of breast cancer in Cyprus is about 45 cases per 100,000 population which is similar to other countries in Southern Europe, Parkin et al. (1999). In several of these countries including Italy, De Benedeti et al. (1998), Turkey, Yazici et al. (2000), Yugoslavia, Papp et al. (1999), and Greece, Ladopoulou et al. (2002) genetic studies in familial breast/ovarian cancer have characterized a number of different mutations in the two breast cancer susceptibility genes. Mutation 5382insC in BRCA1 appears to be the most frequent deleterious mutation in these countries as is the case with most European populations. In

preliminary studies of the Cypriot population, this mutation was not found in breast cancer patients with a moderate to strong family history, Hadjisavvas et al. (2001). Therefore we undertook a more detailed molecular study and this manuscript presents the genetic data, of a comprehensive BRCA1 and BRCA2 mutation analysis in 75 patients from 40 Cypriot families with multiple cases of breast/ovarian cancer.

In the 40 Cypriot families investigated, 4 BRCA deleterious mutations were characterized. One mutation, 5429delG was detected in the BRCA1, and two mutations, C1913X frameshift 8984delG were detected in BRCA2. The BRCA2 8984delG occurred in three unrelated families, so in total, 3 deleterious mutations were found in 5 of the 40 families analysed. This gives a percentage of 12.5% positive families, which is similar to other European high risk populations. It appears that in the Cypriot population, BRCA2 plays a more significant role in the familial breast cancer phenotype since fewer mutations were found in BRCA1, Hadjisavvas et al. (2001, 2003).

Of particular interest is the BRCA2 frameshift mutation, 8984delG, that appears to be a recurrent mutation in our population, as it was detected in 3 unrelated families. Although no population has demonstrated founder effects as striking as those observed in Ashkenazi Jews or of Icelandic origin, haplotype analysis showed that this mutation is a founder mutation in the Cypriot population. This is a new and important finding that has implications for screening families that are at a high risk for developing breast and ovarian cancer. Two of the three deleterious mutations detected are novel, namely the BRCA1 truncating mutation 5429delG and the BRCA2 C1913X, so they could be unique to the Cypriot population.

In conclusion it appears that the BRCA2, plays a more important role than BRCA1 in the familial breast cancer phenotype in the Cypriot population. The identification of a founder mutation in our population has important implications, for screening breast/ovarian cancer within Cyprus.

Acknowledgement

We thank the family members for their willingness to cooperate. We also acknowledge the financial support of the Ministry of Health and the Research Promotion Foundation of Cyprus, through grant 32/2001.

References

1. Antoniou, A.C., Pharoah, P.D.P., McMullan, G., Day, N.E., Stratton, M.R., Peto, J., Ponder, B.J., Easton, D.F. (2002). A comprehensive

model for familial breast cancer incorporating BRCA1, BRCA2 and other genes, British J of Cancer, 86, 76-83.

2. Breast Cancer Information Core Database (BIC) http://www.nchgr.nih.gov//Intramural_research/Lab_transfer/Bic.
3. Colaborative Group on Hormonal Factors in Breast Cancer. (2001). Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58209 women with breast cancer and 101986 women without the disease, Lancet, 358, 1389-1399.
4. De Benedetti, V.M.G., Radice, P., Pasini, B., Stagi, L., Pensotti, V., Mondini, P., Manoukian, S., Conti, A., Spatti, G., Rilke, F., Pierotti, M.A. (1998). Characterization of ten novel and 13 recurring BRCA1 and BRCA2 germline mutations in Italian breast and/or ovarian carcinoma patients, Hum Mutat, 12, 215-9.
5. Dumitrescu, R.G., Cotarla, I. (2005). Understanding breast cancer risk - where do we stand in 2005?, J Cell Mol Med, 9, 208-221.
6. Hadjisavvas, A., Charalambous, E., Adamou, A., Christodoulou, C.G., Kyriacou, K. (2003). BRCA2 germline mutations in Cypriot patients with familial breast / ovarian cancer, Hum Mutat, 21, 171.
7. Hadjisavvas, A., Neuhausen, S.L., Hoffman, M.D., Adamou, A., Newbold, R.F., Kyriacou, K.C., Christodoulou, C.G. (2001) BRCA1 germline mutations in Cypriot breast cancer patients from 26 families with family history, Anticancer Res, 21, 3307-3311.
8. Hodgson, S.V., Haites, N.E., Caligo, M., Chang-Claude, J., Eccles, D., Evans, G., Moller, P., Morrison, P., Steel, C.M., Stoppa-Lyonnet, D., Vasen, H. (2000). A servery of the current clinical facilities for the management of familial cancer in Europe. European Union BIOMED II Demonstration Project: Familial Breast Cancer: audit of a new development in medical practice in European centres, J Med Genet, 37, 605-607.
9. Ingvarsson, S. (2004). Genetics of breast cancer, Drugs of Today, 40, 991-1002.
10. Ladopoulou, A., Kroupis, C., Konstantopoulou, I., Ioannidou-Mouzaka, L., Schofield, A.C., Pantazidis, A., Armaou, S., Tsiagas, I., Lianidou, E., Efstathiou, E., et al. (2002). Germline BRCA1 and BRCA2 mutations in Greek/ovarian cancer families: 5382insC is the most frequent mutation observed, Cancer Letters, 185, 61-70.
11. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W. et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1, Science, 266, 66-71.
12. Neuhausen, S.L. (1999). Ethnic differences in cancer risk resulting from genetic variation, Cancer, 86, 2575-2582.

13. Papp, J., Raicevic, L., Milasin, J., Dimitrijevic, B., Radulovic, S., Olah, E. (1999). Germline mutation analysis of BRCA1 and BRCA2 genes in Yugoslav breast/ovarian cancer families, Oncol Rep, 6, 1435-1438.
14. Parkin, D.M. (2001). Global cancer statistics in the year 2000, The Lancet Oncology, 2, 533-543.
15. Parkin, D.M., Pisani, P., Ferlay, J. (1999). Estimates of the wordwide incidence of 25 major cancers in 1990, Int J Cancer, 80, 827-841.
16. de Sanjose, S., Leone, M., Berez, V., Izquierdo, A., Font, R., Brunet, J.M., Louat, T., Vilardell, L., Borras, J., Viladiu, P., Bosch, F.X., Lenoir G.M., Sinilnikova, O.M. (2003). Prevalence of BRCA1 and BRCA2 germline mutations in young breast cancer patients: a population-based study, Int J Cancer, 106, 588-593.
17. Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M., Timmerman, M.M., Brody, L.C., Tucker, M.A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews, N Eng J of Med, 336, 1401-1408.
18. Szabo, C.I., King, M.C. (1997). Population genetics of BRCA1 and BRCA2, Am J Hum Genet, 60, 1013-20.
19. Thorlacius, S., Olafsdottir, G., Tryggvadottir, L., Neuhausen, S., Jonasson, J.G., Tavtigian, S.V., Tulinius, H., Ogmundsdottir, H.M., Eyfjord, J.E. (1996). A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes, Nat Genet, 13, 117-9.
20. Venkitaraman, A.R. (2002). Cancer susceptibility and the functions of BRCA1 and BRCA2, Cell, 108, 171-182.
21. Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2, Nature, 378, 789-92.
22. Yazici, H., Bitisik, O., Akisik, E., Cabioglu, N., Saip, P., Muslumanoglu, M., Glendon, G., Bengisu, E., Ozbiloen, S., Dincer, M., Turkmen, S., Andrulis, I.L., Dalay, N., Ozcelik, H. (2000). BRCA1 and BRCA2 mutations in Turkish breast/ovarian families and young breast cancer patients, Br J Cancer, 83, 737-742.

58

Identifying High-Risk Subgroups for Smoking Dependency and Alcohol Consumption Among Adolescents: A Classification Tree Analysis

Panagiota Kitsantas

East Carolina University, Department of Mathematics, Austin Building 124 Greenville, NC 27858-4353

Abstract: Classification tree analysis partitions a population or sample into homogeneous groups based on a set of predictors. This technique is utilized in the current study to identify high-risk subgroups for smoking dependency and alcohol consumption among adolescents. The data (N=3,610) were generated from cross-sectional surveys of the Florida Anti-Tobacco Media Evaluation. A classification tree was built based on two types of smokers namely, established (smoking dependent) and situational (less smoking dependent). Another tree model was constructed for alcohol consumption, drinkers versus non-drinkers. The predictor variables included sociodemographic characteristics, peer smoking, social and health risks, tobacco counter-marketing exposure, role modeling, parent-child interaction, parental monitoring, monetary resources and accessibility. The results support the important role of peer influence in smoking and alcohol consumption among adolescents. Knowing where to illegally purchase cigarettes was essential in the classification of both established and situational smokers. Accessibility to purchasing alcohol and parental monitoring were some of the primary predictors in identifying high-risk subgroups for alcohol use. This study demonstrates the use of classification trees in profiling smoking dependency and alcohol consumption in adolescents.

Keywords and phrases: Classification trees, smoking dependency, alcohol consumption, adolescents

58.1 Introduction

Identifying high-risk subgroups of adolescents or young adults who are at an increased risk of smoking dependency or excessive alcohol consumption has been one of the primary goals of numerous studies in the field of substance abuse. The tools of data analysis in these studies have been mainly multiple and logistic regression, Chi-square statistics, structural equation modeling and discriminant function analysis. Regardless of the use of the analytical technique and research design (e.g., longitudinal or cross-sectional), various studies have identified a large number of similar psychosocial risk factors associated with onset and higher-level smoking or alcohol abuse. Although studying risk factors using these methodologies has increased our understanding of substance abuse among young adults, little is known about the interactive nature of risk factors and their ability to define high-risk subgroups of individuals who are at risk of excessive alcohol consumption or smoking uptake/dependency. One method that could uncover the hidden interactive nature of a data set and identify segments of a population that are most likely to engage in risky behaviors is classification and regression trees (CART).

CART analysis (Breiman *et al.,* 1984) has the ability to partition populations or samples into subgroups of subjects that share similar characteristics. This methodology has increasingly been applied to health related fields and clinical settings (Bachur and Harper, 2001) where the goal is to produce an accurate classifier and provide understanding into the predictive structure of the data. The CART methodology systematically investigates the effects of all covariates and describes the manner in which they interact with each other. The resulting model is visualized as a tree with daughter and terminal nodes that have been assigned to a class based on the response variable. Tree models can be evaluated using cross validation or test-sample estimation.

Although it is an explorative technique, it can be used to profile adolescents' smoking dependency levels and alcohol consumption by revealing the interactive nature of various risk factors. In the current study, the interactive nature of various predictor variables in identifying high-risk subgroups for smoking dependency and alcohol use among adolescents is explored using classification trees.

58.2 Methods

58.2.1 Subjects and procedure

The data (N=3,610) were generated from two cross-sectional surveys of the Florida Anti-Tobacco Media Evaluation (FAME), which represents an evaluative component for the Florida Tobacco Pilot Program (FTPP). The FTPP was financed with funds received from the tobacco industry as part of the set-

tlement reached with the State of Florida (Settlement Agreement *et al.,* 1997). The FAME evaluation design involved repeated cross-sectional telephone surveys of youth (between ages 12 and 17) to track and monitor ad and campaign awareness. These telephone surveys were conducted after the informed consent was read to the parents and children and permission was obtained. Specific information regarding sampling procedures, and the reliability and validity of the surveys are available elsewhere (Sly *et al.,* 2000; Sly *et al.,* 2001; Sly *et al.,* 2002). The FAME cross-sections that were utilized in this study included the October 2000 (n=1,810) and the May 2001 (n=1,800) surveys. We selected these two cross-sections because their survey instruments were identical as well as for their close temporal proximity. Comparisons among a wide variety of sample characteristics indicated that no significant differences existed between these two cross-sections.

58.2.2 Measures

The smoking dependency variable was the result of locating natural break points within a matrix cross-referencing the number of days smoked in the past month by the number of cigarettes smoked per day on those days. If they smoked five or more cigarettes per day on six or more days in the past month, they were coded as established smokers. The remaining smokers were coded as situational smokers. The outcome variable for alcohol consisted of those who drank at least twice in the past month, namely drinkers, versus those who have never had a drink or non-drinkers. The predictor variables used in the construction of the smoking dependency and alcohol consumption tree models comprised a combination of factors that have shown a discriminative power in predicting the outcome variables. The selection of independent variables was based upon significant bivariate associations, and a priori considerations of the literature. The selected factors represent nine major variable categories including sociodemographic characteristics, peer smoking, social and health risks, tobacco countermarketing exposure, role modeling, parent-child interaction, parental monitoring, monetary resources and accessibility.

58.2.3 Data analysis

In the current study, two classification trees were built: established versus situational smokers; and drinkers versus non-drinkers. The same set of predictor variables was used to build both of these models. In constructing the classification trees, the Gini Index was used in the splitting process, while misclassification costs for each class were set equal. Cross validation was implemented to evaluate the predictive performance of each classifier. The selection of these criteria were based on the size of the data set (e.g., cross validation is commonly used for smaller samples) and prior theoretical considerations of the analytical

technique. The CART software (Salford Systems, 2000) was employed to build classification trees for each outcome measure.

58.3 Results

58.3.1 Established versus situational smokers model

The established vs. situational smokers model (Figure 58.1) consisted only of established smokers (class = 1, n = 207 or 50.6 percent) and situational smokers (class = 0, n = 202 or 49.4 percent). This model revealed three established smoker subgroups, and another three which were associated with situational smokers. The variable categories, which were important in the construction of this model, included peer smoking, accessibility and social and health risks. One subgroup of established smokers was characterized as having at least one friend who smoked, and considered their identity as smokers unimportant (70.4 percent). Established smoking was also more likely among those who consider their identity as smokers important, knowing where to illegally purchase cigarettes, and not receiving an allowance. Receiving an allowance, however, was associated with situational smoking (58.5 percent). Situational smokers were also classified as having no best friends who smoked, and reportedly wearing their seat belt "most of the time" (80.6 percent).

58.3.2 Drinkers versus non-drinkers

The tree model in Figure 58.2 shows a classifier built for drinkers (class = 1, n = 615 or 80.8 percent) versus non-drinkers (class = 0, n = 2594 or 19.2 percent). The variable categories which were important in the construction of this model included peer smoking, accessibility (knowing anyone with a fake ID or a store that sells alcohol), and parental monitoring. Drinkers were characterized as having friends who smoke, are aware of a store that sells alcohol and did not think that friends' parents would tell on them if they saw them smoking.

58.4 Conclusion

In this study, we employed classification trees to identify high-risk subgroups for smoking dependency and alcohol consumption among adolescents. The results support the important role of peer influence and accessibility (knowing where to illegally purchase cigarettes or alcohol) in smoking and alcohol use. Research evidence indicates that peer, parent and family influences are of major importance in stages of smoking greater than experimental (Jackson and Henriksen, 1997). The repetition of variables across the two models also suggests that one should expect to identify similar characteristics in profiling both smoking

dependency and alcohol use levels. This indicates that smoking cessation programs should consider accessibility and peer influence factors at higher levels of smoking dependency and alcohol use.

References

1. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman and Hall, Florida.

2. Bachur, R. G. and Harper M. B. (2001). Predictive model for serious bacterial infections among infants younger than 3 months of age, *Pediatrics*, **108**, 311-316.

3. Jackson, C., and Henriksen, L. (1997). Do as I say: parent smoking, antismoking, socialization, and smoking onset among children, *Addictive Behaviors*, **22**, 107-114.

4. Settlement Agreement, The State of Florida, The American Tobacco Co, et al. (1997). Civil Action No. 95-1466 AH, Fla. Cir., Palm Beach Co., 25 August 1997.

5. Sly, D. F., Heald, G., Hopkins, R. S., Moore, T. W., McCloskey, M. and Ray, S. (2000). The industry manipulation attitudes of smokers and nonsmokers, *Journal of Public Health Management and Practice*, **6**, 49-156.

6. Sly D. F., Hopkins R. S., Trapido E, and Ray, S. (2001). nfluence of a counteradvertising media campaign on initiation of smoking: the Florida "truth" campaign, *Journal of Public Health*, **91**, 233-238.

7. Sly, D. F., Trapido, E., and Ray, S. (2002). Evidence of the dose effects of an antitobacco counteradvertising campaign, *Preventive Medicine*, **35**, 511-518.

8. Salford Systems (2000). CART software Release 4.0 edition, San Diego, California.

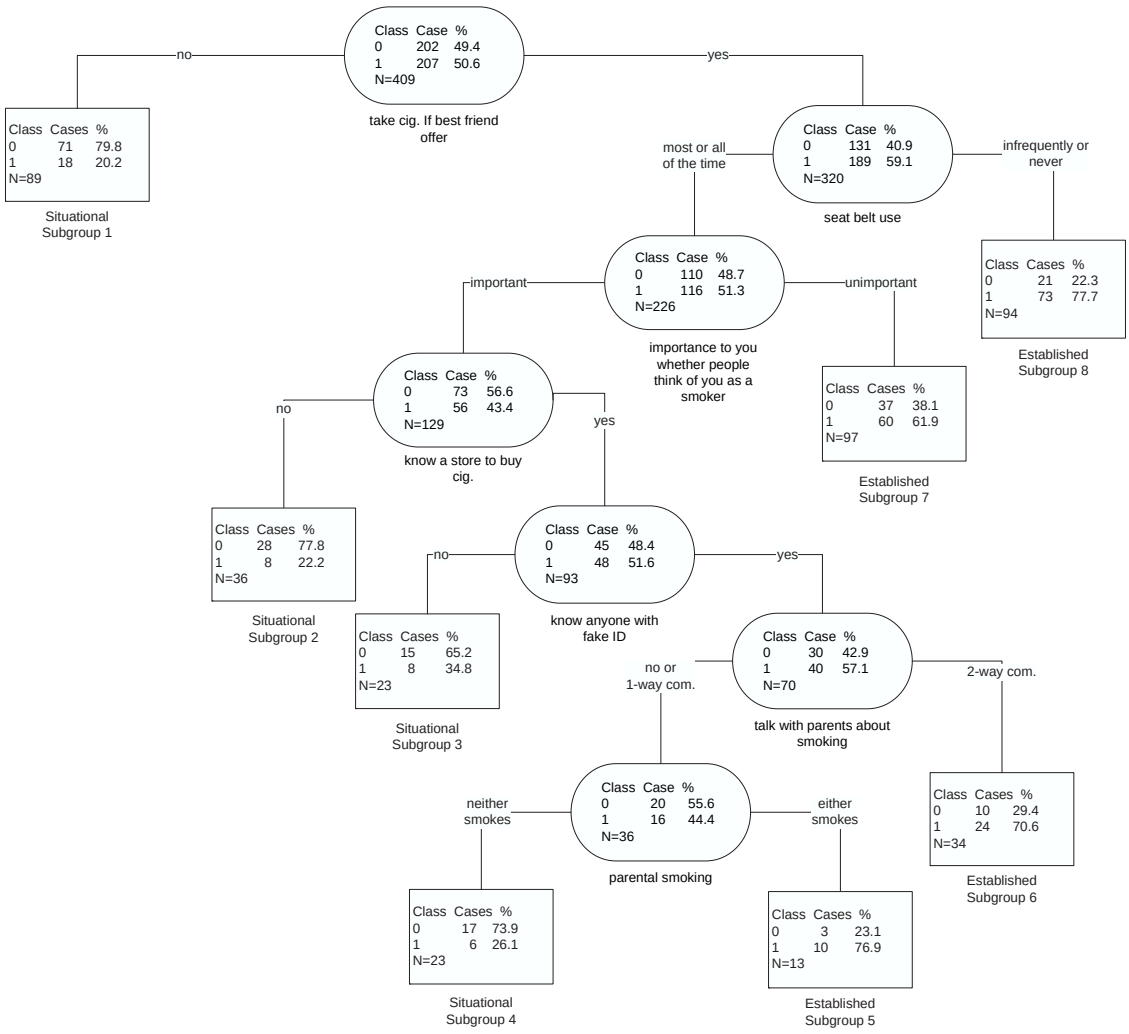


Figure 58.1: Established vs. Situational Smoker Model (class 0 = situational smokers, class 1 = established smokers)

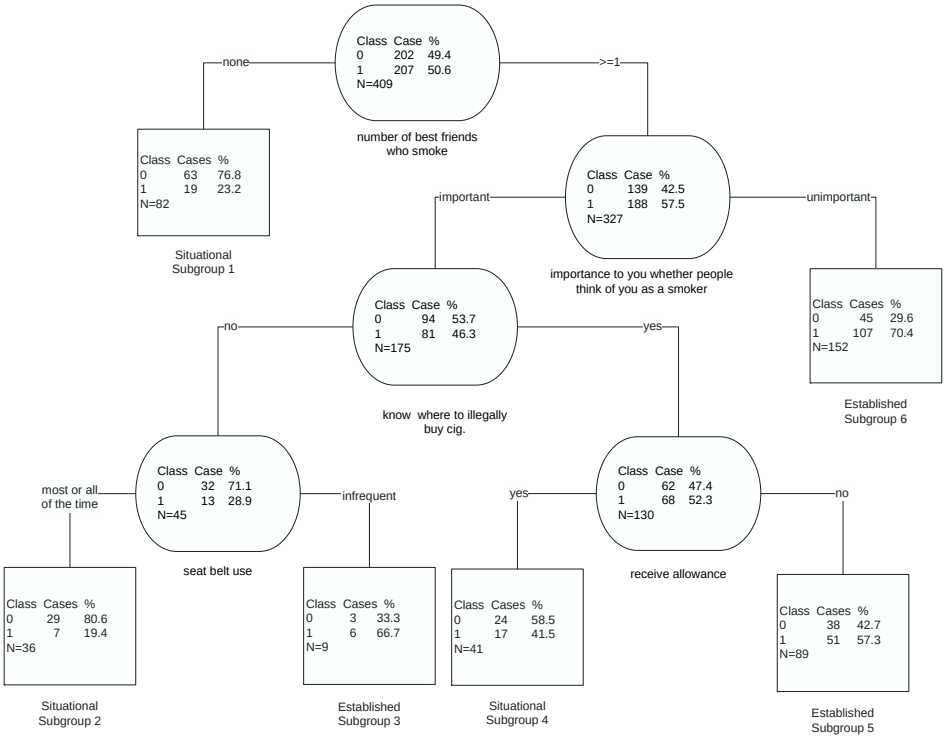


Figure 58.2: Drinkers versus Non-Drinkers (class 0 = non-drinkers, class 1 = drinkers)

59

On The Logit Methods for Ca Problems

Christos P. Kitsos

Department of Mathematics
Technological Institute of Athens

Abstract: The target of this paper is to discuss the Logit model and the Logistic regression, when are applied on data set related to cancer (Ca). The sequential principle is adopted. The paper is part of the research results we are working on the last years.

Keywords and phrases: Logit model, logistic regression, optimal design, canonical form

59.1 Introduction

We study the case, when performing an experiment, the response, Y say, has two outcomes, usually denoted 0 1. Such a response is known as binary response, and is linked with the explanatory variable X, through a probabilistic model $T(x;\theta)$ of the form:

$$T(x;\theta) = P\left(Y = 1 \,|\, x\right) = 1 - P\left(Y = 0 \,|\, x\right) \tag{59.1.1}$$

with θ being the vector of involved parameters and x a value of X. Typical examples in Cancer Bioassays, Kitsos (2005), for $T(x;\theta)$ are the Logit and Probit models, and the explanatory variable, denoting exposure to risk, to be binary i.e. $x = 0$ or 1. For discussion on such cases see Breslow and Day (1980). Usually we denote by $p(x) = T(x;\theta) = P(Y = 1 \,|\, x)$, the conditional probability that a person suffers from Ca, with x being the value of the exposure. Then, a well known, 2×2 contigency table is obtained, Kotti and Rigas (2005) for the Logistic Regression model of the form:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \equiv T(x;\beta_0,\beta_1) \tag{59.1.2}$$

Model (59.1.2) is within the class of the multistage models, Kitsos (1999) and has been investigated under the sequential principle of designing, when the 100%

percentile, L_p say, is to be estimated. In this paper some theoretical results are provided in Sections 2 and 3, while an application is, briefly, discussed in Section 4.

We are refering to Logit Methods to cover both Logit model and Logistic Regression.

59.2 Optimal Designs For The Logit Model

Notice that the Logit Model (59.1.2) is of the form $P(Y = 1 \mid x) = T(\theta^T u)$, with $\theta = (\beta_0, \beta_1)$, $u = (1, x)$. For the parameter $\theta = (\beta_0, \beta_1) \in \mathbb{R}^2$ and ξ a design measure from a family of design measures Ξ the following is true:

Proposition 59.2.1 *For the logistic model as in (59.1.2) Fisher's information matrix $I(\theta, \xi)$, is of the form*

$$I(\theta, \xi) = T'^2[T(1-T)]uu^t \tag{59.2.3}$$

with u^t the transpose of u and T' the derivative of T.

Proof. If we let $z = \theta^t u$ then the log-likelihood function is of the form

$$\ell = \log\{T(z)^Y[-T(z)]^{1-Y} + \text{const}$$

Therefore

$$I(\theta, \xi) = E\{(\nabla\ell)(\nabla\ell)^t\} = T'^2[T(1-T)]uu^t$$

■

With $T(x; \theta)$ the probit model (59.2.3) is still true. For the D-optimal design the design measure is $\xi = 1/2$ and the optimal points are of the form $(\pm 1.54 - \beta_0)/\beta_1$, provided that belong to the design space, otherwise are the endpoints of the design space $S \subset \mathbb{R}$, see Kitsos (1986) for details. For the Logit model, the "canonical form" has been introduced, Kitsos (1986), Ford et al. (1992). This is based on the fact that c-optimality remains invariant, when the data is transformed by

$$g = \begin{pmatrix} 1 & 0 \\ \beta_0 & \beta_1 \end{pmatrix} \tag{59.2.4}$$

That is considering the transformation of the vector u as $w = gu$, under c-optimality with the ray c defined as $c = (1, x)^t$ then it can be proved, Kitsos (1986).

Proposition 59.2.2 *Under c-optimality, with M_i, $i = u, w$ as above, being the corresponding average per observation information matrices in u and z coordinate system then*

$$c^t M_u^{-1} c = c_w^t M_w^{-1} c_w \tag{59.2.5}$$

with $c_w = gc^t$. ■

Now, for this affine transformation g we prove the following

Proposition 59.2.3 *The set of the (affine) transformations G as*

$$G = \left\{ g = \begin{pmatrix} 1 & 0 \\ \beta_0 & \beta_1 \end{pmatrix}, \quad \beta_0, \beta_1 \in \mathbb{R} \right\} \tag{59.2.6}$$

forms a group, under matrix multiplication.

Proof. Considering $h \in G$ as

$$h = \begin{pmatrix} 1 & 0 \\ \alpha_0 & \alpha_1 \end{pmatrix}, \quad \alpha_0, \alpha_1 \in \mathbb{R}$$

then it can be proved easily that

$$hg = \begin{pmatrix} 1 & 0 \\ \gamma_0 & \gamma_1 \end{pmatrix} \in G$$

with $\gamma_0 = \alpha_0 + \alpha_1\beta_0 \in \mathbb{R}$, $\gamma_1 = \alpha_1\beta_1 \in \mathbb{R}$. Similarly if $k \in G$ then it can be proved $(hg)k = h(gk)$, the unit transformation i is the identity matrix, the inverse transformation $g^{-1} \in G$ is the inverse matrix of g and $gi = ig = g$. ■

This theoretical result practically means that we can work as follows: perform the experiment with the "easiest" scale and position parameters. Performing the experiment at the optimal design points as in c-optimality a prior knowledge on β_0, β_1 is needed. Then, transfer the results with an element within the group of affine transformations and still you have an optimal design. Go back to the initial etc. still you are moving within a class of optimal designs. That is perform the remaining experiments within the class of designs generated by the group of transformations G.

From the above discussion for the Logit model and link function defined by the Logit transformation the D-optimal design allocates half observations (i.e. $\xi = 1/2$) at the optimal design points ± 1.5434, when the vector of coefficients in $\theta = (0, 1)$. Different c-optimal designs can be produced with different $\theta = (\beta_0, \beta_1)$ and "direction ray" c. For $\theta = (1, 1)$, $c = (1, 3)$ and the explanatory variable within $[-3, 3]$ the c-optimal design is a two point design at -3 and 1.4164, with the corresponding weights to be equal to 0.1826 and 0.8174 respectivelly. With the group of affine transformations the experimenter can move to different "orbit", adopting the gc^t "direction ray", when he had already evaluated only one set of experiments. That is he saves experiments, when the appropriate theory discussed already is adopted.

59.3 Sequential Logit Methods

As it has been pointed out by Breslow and Day (1980), among others, under the Logit Model the 2×2 contigency table is related to the relative risks, as

well as to the odds ratio. Actually the relative risk is the odd ratio relative to the binary value of x, see Section 1. The involved $\mathcal{X}^2$ test is equivalent on test proportions, that is relative risks. That is why we propose the sequential principle of design to test proportions. The technique can be adopted from the early work of Cox (1963), for testing the odds-ratio or log-odds as

$$\ln \psi = \ln \frac{P_1 Q_0}{P_0 Q_1} \tag{59.3.7}$$

with

$$P_1 = \frac{B_0 B_1}{1 + B_0 B_1}, \qquad P_0 = \frac{B_0}{1 + B_0}$$

$B_0 = \exp(\beta_0)$, $B_1 = \exp(\beta_1)$, $Q_i = 1 - P_i$, $i = 0, 1$ see also Breslow and Day (1980, p. 194) for the notation and Ghosh (1970, p. 363) for a brief discription of the method. Either batch sequential or fully sequential Logistic Regression Methods can be adopted if the logit transform is used to regress the explanatory variable ie

$$y = \text{logit}P = \ln \frac{P}{1 - P} = \beta_0 + \beta_1 x \tag{59.3.8}$$

The parameters can be interpeted as relative risks, Breslow and Day (1980, p. 194), while a sequential design can be applied for the regression model (59.3.8), with the main idea being in Ghosh (1970, p. 359). We are working on this and some results on the augmentation of the relative risks are available.

We believe that Logit Methods can be applied with the sequential principle. In the sequence an example of the classical Logist Regression method is briefly discussed, while a sequential approach is under investigation, as well as the biological interpretation. Moreover the sequential character of the model can be discussed on the basis that the k regressors in the model are increased to $k+1$. Then we can prove that the relative risk of the $k+1$ regressors, RR_{k+1}, is a function of RR_k, the relative risk of k regressors. We believe that this result influences the risk analysis for the multiple Logistic Regression.

59.4 Logistic Regression Analysis for Ca

Various studies have been carried out to investigate the genetic polymorphisms, metabolizing enzymes and risk of Ca patients, see Rossi et al. (2003) among others. On a data set communicated with experts the Logistic Regression fit was crucial: to investigate the age of a woman, the weight, the beginning and end of her period, the age of first child, CYP17(A1/A1, A1/A2, A2/A2), COMT(G/A, A/A, G/G), ER(PP, pp, pP) plus other biological characteristics. Emphasis was given on CYP17 and COMT polymorphisms. The data were collected under Dr. Voutsinas supervision in the area of Patras, Greece, in the year 2004. They include 51 breast Cancer cases and 66 controls. As far as CYP17 concerns

there were 17 cases of type A1/A1, 29 cases of type A1/A2 and 5 cases of type A2/A2. As far as COMT concerns, there were 10 cases of A/A, 30 cases of G/A and 11 cases of G/G. The age of the women that had a baby born was also recorded among other characterestics. We are not going to produce the results here and provide a full analysis. Most of the results on the biological references, as Rosi et al. (2003) cover some points of statistical analysis. But what if the data set increased sequentially, or even by batches. How much do we gain on information? Practically at each stage we can evaluate Fisher's information. But we do not know how "better" are the relative risks evaluated.

This paper tries to cover such questions, or rather to offer new lines of thought on such questions. We also focus on the case when k explanatory variables are needed. We are working on it, evaluating how the log-odds is influenced when the $k+1$ variable enters the k-variable logistic model.

Acknowledgements

I would like to thank Dr M. Voutsinas who brought the data set to my attention and Prof. A. Rigas for helpful discussions. My thanks are extended to the referee for his helpful comments.

References

1. Breslow, N. E. and Day, N. E. (1980). *Statistical Methods on Cancer Research*. In Proc. IARC, Lyon, France.

2. Cox, D. R. (1963). Large Sample Sequential Tests for Composite Hypotheses, *Sankhya A*, **25**, 5-12.

3. Ford, I. Torsney, B. and Wu, C. F. J. (1992). The use of canonical form in the construction of locally optimal designs for nonlinear problems. *J. R. Stat. Soc. B*, **54**, 569-583.

4. Ghosh, B. K. (1970). *Sequential Tests for Statistical Hypotheses*, Addison-Wesley, Canada.

5. Kitsos, C. P. (1986). *Design and Inference for Nonlinear Problems*, Ph. D. Thesis, University of Glasgow, Glasgow, Scotland, U.K.

6. Kitsos, C. P. (1999). Optimal Designs for Estimating the percentiles of the Risk in Multistage Models of Carcinogenesis, *Biometrical Journal*, **41**, 33-43.

7. Kitsos, C. P. (2005). Optimal Designs for Bioassays in Carcinogenesis. In *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, pp. 267-279 (Eds. Lutz Elder and Christos Kitsos), Wiley, Chichester, U.K.

8. Kotti, V. and Rigas, A. (2005). Logistic Regression Methods and their Implementation. In *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, pp. 355-369 (Eds. Lutz Elder and Christos Kitsos), Wiley, Chichester, U.K.

9. Rossi, L. et al. (2003). Genetic Polymorphisms of Steroid Hormone Metabolizing Enzymes and Risk of Liver Cancer in Hepatitis C-infected Patients, *Journal of Hepatology*, **39**, 564-570.

10. Torsney, B. and Mursati, A. (1993). On the construction of optimal design with applications to binary response and to weighted regression models. In *Proc. 3rd Model Oriented Data Analysis Workshop* (Eds. Mueller W., Wynn H. and Zhigljavsky A.), pp. 37-52, Physica-Verlag, Germany.

60

Application of the Sufficient Empirical Averaging Method for Inventory Control Problem Solving

Eugene Kopytov and Catherine Zhukovskaya

Transport and Telecommunication Institute, Lomonosova 1, LV-1019, Riga, Latvia. E-mail: kopitov@tsi.lv
Riga Technical University, Kalku 1, LV-1658, Riga, Latvia E-mail: kat_zuk@hotmail.com

Abstract:

In this article the so-called Sufficient Empirical Averaging (SEA) method is used for inventory control problems solving. It assumes the existence of the complete sufficient statistics for unknown parameters of the distributions. The application of this method allows getting unbiased estimates with minimum variance.

Keywords and phrases: Inventory control, estimation, sufficient statistic

60.1 Introduction

The basic aspect of operation of any company is connected with the inventory control problems. Any mistakes in the planning of inventory control process result in a considerable decrease of the efficiency of a company's operation and of the quality of customers' service. To describe costs, associated with the positive shortage, we could create some safety stock of goods. On the other hand the supplementary stock increases the holding costs.

Different types of stochastic inventory models are considered by Chopra and Meindl (2001), Ross (1992) etc. In practice it is common for inventory manager to answer on two basic questions: how many to order and when to order. There are many different types of inventory control models, which provide the decision-maker with a satisfactory solution.

Statistical problems of the inventory control are considered very seldom though their application is practically impossible without an estimation of stochastic models. In this article one statistical problem of the inventory control on the finite interval of time $(0, T)$ is considered. The similar problems

were described earlier by Kopytov and Greenglaz (2004). We consider a singe-product inventory control model under the following conditions. Initially there are k items of goods in the warehouse. In the course of time the quantity of goods will be decreased. Demand for goods is described by a recurrent flow of arrived claims. The interarrival time has the exponential distribution with the parameter λ, shifted with the value δ:

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-\delta)}, & x > \delta, \\ 0, & otherwise. \end{cases}$$

Each arrived claim requires a random quantity of goods. We supposed that it has normal distribution with parameters μ, σ.

The current claim will be rejected, if the required quantity of goods is too great, namely, it will be greater then the value of function $rej(k,t)$ where k is the current stock level (quantity of goods in the stock) and t is the current time moment. Obviously, $rej(k,t) \leq k$. The problem consists in an estimation of expectation of numbers of the rejected claims on the interval $(0,T)$. This is our efficiency criterion. Thus parameters λ, μ and σ are unknown, but for them the complete sufficient statistics are fixed. Note, that if we are able to solve the given problem, we can consider a problem of function $rej(k,t)$ definition that optimizes efficiency criterion.

60.2 The Method of Problem Solving

We shall solve this problem by means of the SEA method. The mentioned method Chepurin (1994, 1995, 1999) offered in his works. It can be used when the unknown parameters of distributions admit the complete sufficient statistics. This method is based on the fact, that conditional distributions of the random variables, calculated with fixed values of sufficient statistics, don't depend on the unknown parameters of distributions. So, the necessary random variables could be produced according to their conditional distributions. If the applied sufficient statistics are complete this parametrical method gives unbiased estimators with minimum variance.

If we have the exponential distribution, the sufficient statistics for the parameter λ and the sample $\mathbf{X} = (X_1, X_2, ..., X_n)$ is the sum $A = X_1 + X_2 + ... + X_n$.

The conditional probability density function (p.d.f.) of the sample $\mathbf{X}(\mathbf{s})$ at the point $x = (x_1, x_2, ..., x_n)$ is calculated by the formula:

$$f_{\mathbf{X}}(\mathbf{x}, \lambda | A = s) = \frac{(n-1)!}{s^{n-1}}, \ \forall x_i \geq 0, \ x_1 + x_2 + ... + x_n = s. \qquad (60.2.1)$$

We can see that the conditional p.d.f. of the sample $\mathbf{X}(\mathbf{s})$ doesn't depend on the unknown parameter λ.

Often for generation the conditional distributions "special ways" can be used. E.g., for the exponential case, Engen and Lillegard (1997) suggested the following way. Let the value $\mathbf{s}$ of the $\mathbf{S}_n$ is fixed. Firstly we generate n exponential distributed random variables $X_1^0, X_2^0, ..., X_n^0$ with parameter $\lambda = 1$. Secondly we calculate their sum: $S_n^0 = \sum_{i=1}^n X_i^0$. Then random variables of interest $X_1^*, X_2^*, ..., X_n^*$ may be calculated by the formula $X_i^* = \frac{X_i^0 - s}{S_n^0}$, $i = 1, 2, ..., n$.

If we have normal distributed sample $X_1, X_2, ..., X_n$ with the sufficient statistics $a_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - a_n)^2$, then the conditional p.d.f of the sample elements X_i is calculated by the formula:

$$f_{a_n, S_n^2}(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)}\sqrt{\frac{n}{S_n^2}}\frac{1}{n-1}\left[1 - \frac{n}{(n-1)^2 S_n^{2*}}(x - a_n)^2\right]^{\frac{n}{2}-2}, \quad (60.2.2)$$

$$a_n - \frac{n-1}{\sqrt{n}}\sqrt{S_n^2} < x < a + \frac{n-1}{\sqrt{n}}\sqrt{S_n^2},\ n \geq 3. \quad (60.2.3)$$

60.3 Algorithm of Estimation

We will solve considered inventory problem using of the SEA method. In our case the two different distributions are taking place. Let's describe the procedure of generation of random variables with respect to process of demands for goods. The complete sufficient statistics for the exponential distribution are the sum of the sample elements A and the sample size n_a. Then the simulation of the considered process during the given time T is applied. The time intervals between the neighboring demands for goods are generated according to the appropriate conditional distributions of the given fixed values (A, n_a).

In the beginning of suggested procedure n_a random variables are generated according to the exponential distribution with the parameter λ=1. Let's define them as $X_1^0, X_2^0, ..., X_{n_a}^0$. Then let's calculate their sum

$$A^0 = \sum_{i=1}^{n_a} X_i^0. \quad (60.3.4)$$

The needed values of the intervals $X_1^*, X_2^*, ..., X_{n_a}^*$ between claims for goods are calculated by the formula

$$X_i^* = X_i^0(A - \delta n_a)\frac{1}{A^0} + \delta,\ i = 1, 2, ..., n_a. \quad (60.3.5)$$

Due to the other way, according to the Eq.(60.2.2) - (60.2.3) the required size of goods for sequence of claims $Y_1^*, Y_2^*, ..., Y_n^*$ are generated.

Let's describe the algorithm of the inventory process simulation within one run. Firstly we introduce the following notation: A_i is the moment of the i-th claim for goods

$$A_i = \sum_{j=1}^{i} X_j^*. \tag{60.3.6}$$

Let's define N_a as the number of the next claim for goods, which takes place for the present moment t. Let R define the number of claims, which were rejected for the current run.

The modeling algorithm for one run is the following:

Initial date: the sufficient statistics $A = \sum_{i=1}^{n_a} X_i$ and a_n, S_n^2, which have been calculated on the base of the sample of the size $n_d \geq n_a$, $\delta, K, T, rej(k,t)$, $0 < k < K$, $0 < t < T$.

Output date: R - number of rejected claims till the time moment T.

Step 1: *Generation of arrived claims.*

To generate the sequence $\{A_i\}$ of arrival times by Eq.(60.3.4) - (60.3.6).

Step 2: *Generation of demands values.*

Using Eq.(60.2.2) - (60.2.3) for $n = n_d$ to calculate random variables $Y_1^*, Y_2^*, ..., Y_n^*$ which are demands values.

To take $N_a = 1, R = 0, k = K, t = 0$.

Step 3. If $A_{N_a} > T$ then end,

otherwise:

3.1. To take $t = A_{N_a}$.

3.2. If $Y_{N_a}^* < rej(k,t)$ then $k = k - Y_{N_a}^*$, otherwise $R = R + 1$.

3.3. To take $N_a = N_a + 1$ and go to step 3.

In the end of this run the number of rejected demands R is remembered.

Then we repeat steps $1-3$, r times. According to the obtained results of the runs, the frequencies of different values of rejected claims are calculated and the average number of rejected claims as well. When changing the rejected function we must define a function what gives minimal value of rejection probability.

60.4 Numerical Example

In this article a numerical example is considered. Let input data have the following values $n_a = 10$, $A = 12$, $n = 10$, $a_n = 7$, $S_n^2 = 4$, $\delta = 0.6$.

Three types of rejection functions were considered: the threshold rejection function, where the value of the threshold is constant: $r(k,t) = L$, $0 < L \leq K$; the linear rejection function $r(k,t) = \alpha + \beta k$, and the polylinear rejection function $r(k,t) = \alpha + \beta k + \gamma t + \eta kt$.

Let us consider the following numerical values of the coefficient of the rejected function: $L = 4$, $\alpha = 0.4$, $\beta = 0.3, \gamma = 0.2$, $\eta = 0.1$. The obtained results are presented in the tables 1 - 3. It can be seen that the polylineal rejection function gives the best results. Note, that they admit the minimum variance unbiased estimator.

Table 1

The average number of the rejected claims for the threshold rejection function

	$T = 2$	$T = 3$	$T = 4$	$T = 5$	$T = 6$	$T = 7$	$T = 8$	T=9
$K = 10$	0.26	1.07	1.83	2.60	3.42	4.15	5.01	5.69
$K = 12$	0.25	1.02	1.77	2.58	3.30	4.07	4.85	5.58
$K = 14$	0.24	1.00	1.73	2.57	3.28	4.00	4.77	5.54
$K = 16$	0.23	1.00	1.72	2.56	3.21	3.99	4.75	5.48
$K = 18$	0.22	0.98	1.69	2.51	3.17	3.96	4.74	5.48
$K = 20$	0.19	0.95	1.66	2.41	3.11	3.94	4.73	5.32

Table 2

The average number of the rejected claims for the linear rejection function

	$T = 2$	$T = 3$	$T = 4$	$T = 5$	$T = 6$	$T = 7$	$T = 8$	T=9
$K = 10$	0.30	1.06	1.87	2.63	3.33	4.14	4.96	5.67
$K = 12$	0.27	1.02	1.80	2.52	3.31	4.09	4.79	5.64
$K = 14$	0.25	0.94	1.74	2.48	3.21	3.95	4.68	5.47
$K = 16$	0.19	0.92	1.73	2.44	2.91	3.28	3.86	4.77
$K = 18$	0.00	0.17	0.75	1.54	2.23	3.06	3.79	4.51
$K = 20$	0.00	0.10	0.66	1.38	2.21	2.98	3.76	4.45

Table 3

The average number of the rejected claims for the polylinear rejection function

	$T = 2$	$T = 3$	$T = 4$	$T = 5$	$T = 6$	$T = 7$	$T = 8$	T=9
$K = 10$	0.14	0.60	1.24	1.71	2.64	3.12	3.74	4.54
$K = 12$	0.05	0.02	0.75	1.51	2.01	2.37	3.14	3.76
$K = 14$	0.00	0.12	0.75	1.44	1.73	2.27	2.78	3.67
$K = 16$	0.00	0.10	0.55	1.00	1.57	2.20	2.77	3.65
$K = 18$	0.00	0.05	0.28	0.75	1.25	2.04	2.56	3.42
$K = 20$	0.00	0.01	0.19	0.52	0.83	1.23	1.91	2.59

According to the obtained results we can conclude that the proposed approach allows to find the optimal solutions.

60.5 Conclusion

In this article we considered some statistical problems of the inventory control and described the Sufficient Empirical Averaging method for their solving. Numerical examples have been calculated. The results of our research showed that the proposed method allows the solution of various practical problems of inventory control efficiently.

References

1. Andronov, A., Zhukovska, C. and Chepurin, E.V. (2005). On Application of the Sufficient Empirical Averaging Method to System Simulation, In *Proceedings of the 12th International Conference on Analytical and Stochastic Modelling Techniques and Applications* (Eds., Khalid Al-Begain, Gunter Bolch, Miklos Telek), pp. 144–150, ASMTA 2005, Riga.

2. Chepurin, E.V. (1999). On Analitic-Computer Methods of Statistical Inferences of Small Size Data Samples, In *Proceedings of the International Conference Probabilistic Analysis of Rare Events.* (Eds., V.V. Kalashnikov and A.M. Andronov), pp. 180–194, Riga Aviation University, Riga.

3. Chepurin, E.V. (1995). The Statistical Analysis of the Gauss Data Based on the Sufficient Empirical Averaging Method, In *Proceeding of the Russian University of People's Friendship. Series Applied Mathematics and Informatics*, **N 1**, Moscow, Russian, pp. 112–125.

4. Chepurin, E.V. (1994). The Statistical Methods in Theory of Reliability, *Onbozrenije Prikladnoj i Promishlennoj Matematiki, Ser. Verojatnost i Statistika*, **Vol.1, N 2**, Moscow, Russian, pp. 279–330.

5. Chopra, S. and Meindl, P. (2001). *Supply Chain Management*, Prentice Hall, London.

6. Engen, S. and Lillegrad, M. (1997). *Stochastic Simulations Conditioned of Sufficient Statistics*, Biometrica , **Vol. 84, N 1**, pp. 235–240.

7. Greenglaz, L. and Kopytov, E. (2002). *Inventory Theory with Computer Examples*, TTI, ECH, Riga. (In Russian).

61

Optimal and Universally Optimal Two Treatment Repeated Measurements Designs

Stratis Kounias and Miltiadis Chalikias

Department of Mathematics, University of Athens, Greece

Abstract:[1,2] Optimal Repeated Measurements Designs (RMD(t,n,m)) are studied, for estimating direct and residual effects. Necessary and sufficient conditions are presented and specific designs are given for 2, 3, periods. The model of independent errors is considered with and without direct effect-period interactions. Universally optimal designs, introduced by Kieffer (1975), are also given for 3 periods.

Keywords and phrases: Repeated measurement designs, crossover, changeover, optimal, universally optimal, uniform, balanced, direct effects, residual or carryover effects, clinical trials

61.1 Introduction

Experimental designs in which there are t treatments, n experimental units (e.u.) and an experimental unit is allocated to a sequence of the treatments under investigation over m successive periods, one treatment at the beginning of each period, are called Repeated Measurements Designs, RMD(t, n, m). Other names used are crossover designs or changeover designs, an administered treatment in one period might be administered again in another period.

A direct effect of a treatment, applied at the beginning of a period, is measured at the end of the period, along with the other effects, a carryover effect is the effect of the persistence of the treatment administered in the previous period. We will assume that no treatment effect persists more than one period after its application.

[1]Part of this work was done while the first author was Visiting Professor at the University of Cyprus.

[2]For the second author the project is co-financed within Op. Education by the ESF (European Social Fund) and National Resources.

If Y_{ijk} is the response in the ith sequence of treatments, in period j, on the kth experimental unit, then the model for the continuous response is,

$$Y_{ijk} = \mu+\tau_{ij}+\pi_j+\delta_{i,j-1}+\gamma_i+e_{ijk} \quad j = 1,\ldots,m, \quad i = 0,1,\ldots,2^m-1, \; k = 1,\ldots,n$$

where μ is the general mean, $\tau_{ij} \in \{\tau_A, \tau_B\}$ is the direct effect of the treatment applied in the ith sequence and the jth period, π_j is the effect of the jth period, $\delta_{i,j-1} \in \{\delta_A, \delta_B\}$ is the carryover effect of the treatment applied in the ith sequence and the (j-1)th period, γ_i is the effect of sequence i and e_{ijk} the error, considered continuous random variable with constant variance σ^2 and mean 0.

In this paper we examine the case of t=2 treatments A, B. The observations of a sequence of m treatments, administered on the same unit, are usually dependent, in this work however we study the case of independent observations, within and across units.

The ith sequence effect γ_i could have been taken as the effect γ_{ik} of the kth unit in sequence i, but for the study of optimality of the designs this distinction makes no difference, then we will use .the ith sequence effect γ_i.

The restrictions imposed on the parameters for the uniqueness of the model are: $\tau_{\mathrm{B}} = \pi_m = \gamma_d = 0, \;\; d = 2^m - 1$, no restrictions are imposed on carryover effects because in the first period we do not have carryover effect, i.e. $\delta_{i,0} = 0$.

We are interested in estimating the direct treatment effect τ_A which actually measures the difference $(\tau_A - \tau_B)$ due to the imposed restriction $\tau_B = 0$, we are also interested in estimating the carryover effects $\{\delta_A, \delta_B\}$.

61.2 The Enumeration of Sequences

The dyadic system is used to enumerate the sequences, setting 0 for A and 1 for B, so in 3 periods the sequence AAA is sequence 0, denoted s(0,3), the sequence ABB is s(6,3), in 6 periods the sequence ABBAAB is s(38,6) because $0 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 + 0 \cdot 2^4 + 1 \cdot 2^5 = 38$.
There are 2^m sequences in RMD(t=2,n,m), i.e. $s(0,m), s(1,m), \ldots, s(2^m-1,m)$ and a design is defined by the sequences which are administered.

61.3 Two-Treatment Designs

Model (1) was introduced by Hedayat and Afsarinejad (1975, 1978) who applied the theory of optimal designs to RMD(t,n,m). Kiefer (1975) has given criteria for a design $d^\star$ to be universally optimal, i.e. $\phi(\mathbf{C}_{d*}) \leq \phi(\mathbf{C}_d)$ for all convex functions ϕ with some monotonic properties, then this design will be D,A,E optimal.

Kershner and Federer (1981), give some designs with good properties, of the 6 designs they give, with the names D2.3.3, D 2.3.4, D2.3.5, D4.3.1, D4.3.2,

D4.3.3, D6.3.1, only D 2.3.4 is optimal, the remaining 5 are not optimal. Also the optimal design, in 3 periods, ABB, BAA and for even n, was given by Laska et all (1983), Laska and Meisner (1985), Mathews (1987, 1990), Kushner (1997). This design is universally optimal as has been proved by Cheng and Wu (1980).

Quenouille (1953) has given, in 4 periods, the design ABBA, BAAB, AABB, BBAA for n=0mod 4 with n/4 e.u. to every sequence, this solution has also been given by Laska, Meisner and Kushner (1983) and Mathews (1990), Laska and Meisner (1985).

The least squares estimator for direct effect is given by the relation:

$$(\mathrm{X}_{\mathbf{1}}^{\mathrm{T}}\mathrm{X}_{\mathbf{1}} - \mathrm{X}_{\mathbf{1}}^{\mathrm{T}}\mathrm{P}\mathrm{X}_{\mathbf{1}})\hat{\tau}_{\mathbf{A}} = \mathbf{X_1^T}(\mathbf{I}-\mathbf{P})\mathbf{Y} \tag{2}$$

then

$$\mathrm{var}(\hat{\tau}_{\mathrm{A}}) = \sigma^2 Q^{-1}\,, \quad Q = (\mathrm{X}_{\mathbf{1}}^{\mathrm{T}}\mathrm{X}_{\mathbf{1}} - \mathrm{X}_{\mathbf{1}}^{\mathrm{T}}\mathrm{P}\mathrm{X}_{\mathbf{1}}) = \mathrm{X}_{\mathbf{1}}^{\mathrm{T}}(\mathrm{I}_{\mathbf{4n}} - \mathbf{P})\mathbf{X_1} \tag{3}$$

where $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$, $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, $\mathbf{P} = \mathbf{X_2}(\mathbf{X_2^T X_2})^{-1}\mathbf{X_2^T}$ and $\mathbf{X_1}$ is the column of $\mathbf{X}$ corresponding to the parameter τ_A.

Our task is to find the design minimizing $var(\hat{\tau}_A)$ i.e. maximizing Q, this is called optimal design for estimating τ_A.

We denote by u_i the number of e.u. administered to sequence I and the problem is to find $u_i, i = 0, 1, \ldots, 2^m - 1$ for which Q is maximized.

If PX_1 is the orthogonal projection of X_1 onto the linear space of the columns of X_2, then Q is the square of the X_1 distance of X_1 from PX_1.

We find that Q can be expressed as $Q = \frac{1}{m}\left[R - \mathbf{q^T M^{-1} q}\right]$, where R, $\mathbf{q} : mx1$, $\mathbf{M} : mxm$ are functions of $u_i, i = 0, 1, \ldots, 2^m - 1$ and $\mathbf{M}$ is non negative definite.

61.4 Optimal Designs

A lot of work has been done on optimal RMD, see also Kunert (1983), Jones, B. and Kenward, M.G. (2003).

61.4.1 Two Periods

a) n even: k e.u. to the sequences AA, AB and (n-2k)/2 e.u. to BA, BB. The minimum variance is $\mathrm{var}(\hat{\tau}_{\mathrm{A}} - \hat{\tau}_{\mathrm{B}}) = \sigma^2 8/n$.

b) n odd: (n-1)/2 e.u. to AA and (n+1)/2 e.u. to AB, with $\mathrm{var}(\hat{\tau}_{\mathrm{A}} - \hat{\tau}_{\mathrm{B}}) = \sigma^2 8/(n - \frac{1}{n})$.

If in the model exists the interaction $(\tau\pi)_{A1}$, the we can only estimate $(\tau_A - \tau_B) - (\tau\pi)_{A1}$ and the optimal design is the same as above and with the same variance.

For estimating the carry over effects δ_A, δ_B, we can only estimate $\delta_A - \delta_B$ and the optimal design is.

c) n even: n/2 e.u. to each of the sequences AA, BB with $\text{var}(\hat{\delta}_\text{A} - \hat{\delta}_\text{B}) = \sigma^2 8/n$

d)n odd: (n-1)/2 e.u.. to AA and (n+1)/2 e.u. to BB with $\text{var}(\hat{\delta}_\text{A} - \hat{\delta}_\text{B}) = \sigma^2 8/(n - \frac{1}{n})$.

If the interaction $(\tau\pi)_{A1}$ is in the model, then we can only estimate $\delta_\text{A} - \delta_\text{B} - 2(\tau\,\pi)_\text{A1}$ and the optimal design is the same as that given in (c) and (d).

For estimating both parameters $(\tau_A, (\delta_A - \delta_B))$, the D optimal design is to allocate equal or almost equal e.u. to each of the sequences AA, AB, BA, BB.

61.4.2 Three Periods

For estimating $\tau_A - \tau_B$ or $\delta_A - \delta_B$, the optimal design is the same i.e.,

a) n even: n/2 e.u. to ABB and BAA with $\text{var}(\hat{\tau}_\text{A} - \hat{\tau}_\text{B}) = \sigma^2 3/(2n)$, and $\text{var}(\hat{\delta}_\text{A} - \hat{\delta}_\text{B}) = \sigma^2 2/n$

b) n odd: (n+1)/2 e.u. to ABB and (n-1)/2 e.u. to BAA, with variance $\text{var}(\hat{\tau}_\text{A} - \hat{\tau}_\text{B}) = \sigma^2(3n)/(2(n^2-1))$ and $\text{var}(\hat{\delta}_\text{A} - \hat{\delta}_\text{B}) = \sigma^2(2n)/(n^2-1)$.
The above design is universally optimal.

61.4.3 Four Periods

The optimal designs have 2 A and 2 B, we only mention that the optimal designs for estimating direct effects are $\text{var}(\hat{\tau}_\text{A} - \hat{\tau}_\text{B}) = \sigma^2(Q^\star)^{-1}$, where,

$$1) \qquad n = 0 mod 4 : Q^\star = n,$$

$$2) \qquad n = 2 mod 4 : Q^\star = n - \frac{4}{11n},$$

$$3) \qquad n = 1 mod 4 : Q^* = n - \frac{2(6n+5)}{11n(n+1)},$$

$$4) \qquad n = 3 mod 4 =: Q^* = n - \frac{2(6n^2-3n-1)}{11n(n^2-1)}.$$

61.5 Universally Optimal Designs

In the previous sections we have given optimal designs for estimating either direct effects i.e. $(\tau_A - \tau_B)$ or residual effects i.e. $(\delta_A - \delta_B)$.

In special cases when the number of e.u. is n=0mod 2 or mod 4, there exist universally optimal designs for estimating both parameters i.e. $(\tau_A - \tau_B)$, $(\delta_A - \delta_B)$.

If a design is universally optimal then it is D, A, E optimal, Kiefer (1975).

Cheng and Wu (1980), Theorem 3.1 and Theorem 3.3, give criteria for universally optimal designs which are uniform and strongly balanced and they refer to model (1).

The same authors show, Theorem 3.4, that universally optimal designs (i) are also optimal for estimating $(\tau_A - \tau_B)$ and in Theorem 3.5 they show that universally optimal designs (ii) are optimal for estimating $(\delta_A - \delta_B)$.

We show that in both cases, universally optimal designs are optimal for estimating either $(\tau_A - \tau_B)$ or $(\delta_A - \delta_B)$. No designs exist that satisfy Theorems 3.1 and 3.2 of Cheng and Wu (1980) in two periods.

61.5.1 Three Periods

The only design which is uniform in the units in the first m-1=2 periods is (AB, BA). In the 3 periods we have the design which is form by the sequences $s(1,3) = BAA,\ s(2,3) = ABA,\ s(5,3) = BAB,\ s(6,3) = ABB$.

If u_1, u_2, u_5, u_6 is the number of e.u. allocated to each one of the above sequences, then the total number of appearances of the pairs AA, BA, AB, BB is:$AA \to u_1 + u_3$, $BA \to u_1 + u_2 + u_5$, $AB \to u_2 + u_5 + u_6$, $BB \to u_6$. Therefore, to have equality the following holds: $u_1 = u_1 + u_3 + u_5 = u_2 + u_5 + u_6 = u_6 \Leftrightarrow u_3 = u_5 = 0, u_1 = u_6$ and the only universally optimal design is $d^\star = \{BAA, ABB\}$ with equal number of e.u. to each sequence, then n=0mod 2. This design is universally optimal and $\text{var}(\hat{\tau}_\text{A} - \hat{\tau}_\text{B}) = \sigma^2 \frac{3}{2n}$, $\text{var}(\hat{\delta}_\text{A} - \hat{\delta}_\text{B}) = \sigma^2 \frac{2}{n}$.

References

1. Cheng, C.S. and C.F. Wu (1980). Balanced repeated measurements designs. *Ann. Statist.* 11, 29-50. Correction (1983) 11, 349.

2. Hedayat, A. S.and Afsarinejad K.(1975). Repeated measurements designs I. *A survey of Statistical Designs and Linear Models*, J.N. Srivastava editor, North-Holland, Amsterdam, 229-242.

3. Hedayat, A. S. and Afsarinejad K.(1978). Repeated measurements designs II. *Ann. Statist.* 18, 1805-1816.

4. John, J.A., Quenuille, M.H. (1977). *Experiments: Design and Analysis.* Charles Griffin & Co. Ltd, London.

5. Jones, B. and Kenward, M.G. (2003). *Design and Analysis of Cross-Over Trials.* Chapman & Hall/CRC

6. Kiefer, J. (1975). Construction and optimality of generalized Youden squares. *A survey of Statistical Design and Linear Models.* J.N. Srivastava, Editor, North-Holland, Amsterdam, 333-353.

7. Kershner, R.P.and Federer, W.T. (1981) Two-treatment Crossover Design for Estimating a variety of Effects. *J. Am. Statist. Assoc.*, vol 76, 612-619.

8. Kunert, J. (1983). Optimal design and refinement of the linear model with applications to repeated measurements designs. *Ann. Statist.*, 11, 247-257.

9. Kushner, H.B. (1997). Optimality and Efficiency of the Two Treatment Repeated Measurements Design. *Biometrika*, 84, 455-468.

10. Laska, E.M., Meisner M. and Kushner H.B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics*, 39, 1087-1091.

11. Laska, E.M. and Meisner M. (1985). A variational approach to optimal two treatment crossover designs: application to carryover effect models. *J. Am. Statist. Assoc.*, 80, 704-710.

12. Mathews, J.N.S. (1987). Optimal crossover designs for the comparison of two treatments in the presence of carryover effects and autocorrelated errors. *Biometrika*, 74, 2, 311-320.

13. Mathews, J.N.S. (1990). Optimal dual-balanced two treatment crossover designs. *Sankhya*, 52, B, 3, 332-337.

14. Mathews, J.N.S. (1994). Modeling and optimality in the design of crossover studies for medical applications. *Journal of Statistical Planning and Inference*, 42, 89-108.

62

Generalized Linear Models for Marked Point Processes

David Kraus

Institute of Information Theory and Automation
Pod Vodárenskou věží 4
CZ-182 08 Praha 8
Czech Republic

Abstract: When we observe random variables (marks) at random time points, the data can be viewed as realisations of a marked point process. This paper deals with generalized linear models describing the dependence of the mark distribution on predictable covariates.

Keywords and phrases: Marked point process, generalized linear model, iteratively reweighted least squares, smoothing, time-varying regression effects

62.1 Introduction

Consider n individuals observed over the time period $[0, \tau]$. Each of these n subjects experiences a finite number of events occuring at random times

$$0 < T_{i,1} < T_{i,2} < \cdots < \tau.$$

At each event time $T_{i,k}$ of the i-th individual, a random variable $Z_{i,k}$, called mark, is observed. Hence, the i-th observation consists of the pairs

$$(T_{i,1}, Z_{i,1}), (T_{i,2}, Z_{i,2}), \ldots \tag{62.1.1}$$

These pairs form a marked point process (MPP) whose behaviour is driven by the intensity of the event time points and the conditional distribution of the marks.

In addition to the pairs $(T_{i,k}, Z_{i,k})$, we observe some covariates that serve as explanatory variables in regression models for the conditional mark distribution and for the intensity. Our main aim is to study the mark distribution which is modeled by generalized linear models (GLM). We extend results of Martinussen and Scheike (2001) who studied linear models.

The structure of the paper is following. Section 62.2 provides a summary of the stochastic structure of marked point processes. In Section 62.3, regression models for the mark distribution and time-process intensity are formulated. Section 62.4 is devoted to the presentation of the estimation procedure. In Section 62.5, the contribution is closed by comments on future plans.

62.2 Marked point processes

The set of the pairs (62.1.1) is called a marked point process. The time points belong to the interval $[0, \tau]$, the marks take values in a mark space $(E, \mathcal{E})$. It is advantageous to view a marked point process as a measure. Following Brémaud (1981, Chapter VIII), we denote the MPP (62.1.1) by $p_i(dt \times dz_i)$. The object $p_i(dt \times dz_i)$ is a random counting measure on the product $[0, \tau] \times E$. The process

$$N_i(t, A) = \int_0^t p_i(ds \times A) = \int_0^t \int_A p_i(ds \times dz_i)$$

is a counting process (with respect to a filtration, say $(\mathcal{F}_t, t \in [0, \tau])$) which counts every event whose mark lies in $A \in \mathcal{E}$.

We say that $p_i(dt \times dz_i)$ admits the intensity kernel $\lambda_i(t, dz_i)$, if for each $A \in \mathcal{E}$ the counting process $N_i(t, A)$ has the intensity

$$\lambda_i(t, A) = \int_A \lambda_i(t, dz_i).$$

We will suppose that the intensity can be written in the form

$$\lambda_i(t, dz_i) = \lambda_i(t)\Phi_i(t, dz_i),$$

where $\lambda_i(t) = \lambda_i(t, E)$ is the intensity of the process $N_i(t, E)$ counting all the event points regardless of their marks, and $\Phi_i(t, dz_i)$ is the conditional mark distribution given the history up to t and given t is an event point.

62.3 Regression models

In the following, it is assumed that both components of the MPP depend on a set of covariates. Hence, the conditional mark distribution $\Phi_i(t, dz_i)$ as well as the intensity $\lambda_i(t)$ are described by regression models. We consider the following model form.

62.3.1 Generalized linear models for marks

Assume that the mark distribution $\Phi_i(t, dz_i)$ follows a generalized linear model (McCullagh and Nelder, 1989). This means that its expectation $\mu_i(t)$ depends

on the linear predictor $\eta_i(t)$ through the link function g by

$$g(\mu_i(t)) = \eta_i(t) = X_i(t)^\mathsf{T}\beta(t).$$

Here $X_i(t)$ is a p-vector of (possibly time-dependent predictable) covariates and $\beta(t)$ is a vector of time-varying regression coefficients which are modeled non-parametrically. Variance of $\Phi_i(t, dz_i)$ equals $\psi(t)V(\mu_i(t))$, where the dispersion parameter $\psi(t)$ is allowed to depend on time.

For example, for the Poisson regression, the natural (canonical) link is $g(\mu) = \log\mu$, the variance function is $V(\mu) = \mu$ and the dispersion parameter is $\psi \equiv 1$. For the Gaussian linear model studied by Martinussen and Scheike (2001), we have $g(\mu) = \mu$, $V(\mu) = 1$ and $\psi(t) = \sigma^2(t)$.

62.3.2 Aalen's regression for the time process

Similarly to Martinussen and Scheike (2001), the intensity $\lambda_i(t)$ of the counting process $p_i(dt \times E)$ is supposed to follow the Aalen additive regression model

$$\lambda_i(t) = Y_i(t)\alpha(t)^\mathsf{T}U_i(t),$$

where $U_i(t)$ is an r-vector of covariates whose effects are $\alpha(t)$ and $Y_i(t)$ is the risk indicator process. The Aalen model is easy to estimate yet flexible enough for our purpose, as our main goal is to study the distribution of marks.

62.4 Estimation

We will estimate the cumulative regression coefficients $B(t) = \int_0^t \beta(s)ds$, $t \in [0, \tau]$, by a piecewise constant estimator with jumps at the event times. The estimation of its increments is based on an estimating equation. The reason for estimating $B(t)$ rather than $\beta(t)$ is that we wish to make inference about the whole regression functions.

62.4.1 Estimating equation

The estimating equation is

$$\sum_{i=1}^{n} \frac{X_i(t)}{g'(\mu_i(t))\psi(t)V(\mu_i(t))} \left[\int_E z_i p_i(dt \times dz_i) - \hat{\lambda}_i(t)\mu_i(t)dt\right] = 0, \qquad (62.4.2)$$

where $\hat{\lambda}_i(t)$ is an estimate of the intensity $\lambda_i(t)$ of the i-th counting process (based on kernel smoothing of standard estimates of the Aalen model). This equation can be justified from the quasi-likelihood point of view because

$$\frac{X_i(t)\lambda_i(t)}{g'(\mu_i(t))} = \frac{\partial}{\partial\beta(t)} \mathsf{E}\left[\int_E z_i p_i(dt \times dz_i)\Big|\mathcal{F}_{t-}\right] = \frac{\partial}{\partial\beta(t)}\lambda_i(t)\mu_i(t)dt$$

and

$$\psi(t)V(\mu_i(t))\lambda_i(t)dt = \operatorname{var}\left[\int_E z_i p_i(dt \times dz_i)\Big|\mathcal{F}_{t-}\right].$$

It is seen that the dispersion parameter $\psi(t)$ can be cancelled in the estimating equation (62.4.2) and the problem can be solved without knowledge of $\psi(t)$.

62.4.2 Algorithm: IRLS with smoothing

The estimation procedure is a combination of the Iteratively Reweighted Least Squares (IRLS) algorithm and kernel smoothing between steps of the IRLS.

The idea of smoothing between iteration steps is inspired by Martinussen *et al.* (2002) who use the Newton–Raphson algorithm with smoothing for the Cox model with time-varying coefficients. We need to smooth estimates of the cumulative coefficients because terms like $\mu_i(t)$ entering in the estimating equation (62.4.2) are evaluated in the original (noncumulative) functions. Another reason for smoothing is following. At any time, at most one individual experiences event, and, thus, the information is limited. Therefore, it is impossible to iterate at each event time separately. Instead, Martinussen *et al.* (2002) propose to use smoothing in order to stabilize the procedure.

Having a piecewise constant function $\tilde{B}(t)$ (the previous iteration), we use the kernel smoother of the form

$$\tilde{\beta}(t) = \int_0^\tau \frac{1}{b_\beta} K\left(\frac{s-t}{b_\beta}\right) d\tilde{B}(s),$$

where K is a zero-mean unit-variance kernel supported on $[-1,1]$ (e.g., Epanechnikov) and b_β is a bandwidth parameter.

Now we can describe the estimation algorithm. Denote the previous iteration $\tilde{B}(t)$. The iterations go as follows:

(1) Smooth $\tilde{B}(t)$ to obtain $\tilde{\beta}(t)$.

(2) Perform one step of the IRLS for all the event times $t \in [0,\tau]$:

(2i) Compute the 'working response'

$$\tilde{r}_i(t)dt = \tilde{\eta}_i(t)dt + \frac{g'(\tilde{\mu}_i(t))}{\hat{\lambda}_i(t)}\left[\int_E z_i p_i(dt \times dz_i) - \hat{\lambda}_i(t)\tilde{\mu}_i(t)dt\right].$$

(Here we denote by a tilde quantities with $\beta(t)$ replaced by $\tilde{\beta}(t)$.)

(2ii) Compute the weights

$$W_i(t) = \frac{\hat{\lambda}_i(t)}{g'(\tilde{\mu}_i(t))^2 V(\tilde{\mu}_i(t))}.$$

Set $W(t) = \operatorname{diag}[W_i(t)]$.

(2iii) Regress $\tilde{r}(t)dt$ on $X(t)$: obtain the new iteration $d\tilde{B}(t)$ by the weighted least squares

$$\begin{aligned} d\tilde{B}(t) &= \left[X(t)^{\mathsf{T}} W(t) X(t)\right]^{-1} X(t)^{\mathsf{T}} W(t) \tilde{r}(t) dt \\ &= \tilde{\beta}(t)dt + \left[X(t)^{\mathsf{T}} W(t) X(t)\right]^{-1} X(t)^{\mathsf{T}} W(t) \\ &\quad \times \operatorname{diag}\left[\frac{g'(\tilde{\mu}_i(t))}{\hat{\lambda}_i(t)}\right]\left[\int_E zp(dt \times dz) - \hat{\lambda}(t)\tilde{\mu}(t)dt\right]. \end{aligned}$$

($\hat{\lambda}(t)\tilde{\mu}(t)$ is a componentwise product of the two vectors.)

(3) Go to (1).

Of course, the algorithm requires some initial values. For the first iteration, we replace (1) by computing a locally polynomial estimate of $\beta(t)$.

Our experience based on simulations shows that the number of iterations about 15 is usually enough (often, even after less than 10 iterations the procedure stabilizes).

Note that when the link is canonical (i.e., $g'(\mu)V(\mu) = 1$), the IRLS algorithm coincides with the Newton–Raphson algorithm.

62.5 Final comments

Simulations show that the described procedure yields consistent estimates. Theoretic results on consistency and asymptotic distribution of the estimates are in preparation.

Further topics of our interest in this kind of models include development of tests of time-constancy of the regression functions and study of semiparametric models with some of the coefficients constant and some time-varying.

Acknowledgement

The author gratefully acknowledges the support from the Czech Republic Grant GAČR 201/05/H007.

References

1. Brémaud, P. (1981). *Point Processes and Queues. Martingale Dynamics.* Springer, New York.

2. Martinussen, T. and Scheike, T. H. (2001). Sampling adjusted analysis of dynamic additive regression models for longitudinal data. *Scandinavian Journal of Statistics*, **28**, 303–323.

3. Martinussen, T., Scheike, T. H. and Skovgaard, I. M. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scandinavian Journal of Statistics*, **29**, 57–74.

4. McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall, London–New York.

63

A Comparison of Discriminant Analysis and Logistic Regression for the Prediction of In-hospital Mortality Among Patients Hospitalized With A Range Spectrum of Acute Coronary Syndromes

Georgia Kurlaba, Demosthenes B. Panagiotakos

Department of Nutrition and Dietetics, Harokopio University, Athens, Greece

63.1 Introduction

Both discriminant analysis and logistic regression can be used to predict the probability of a specified categorical outcome using several available variables.

The primary objective of this study was to investigate if these two methods of analysis result in the same patients' characteristics that are indicative of patients admitting with Acute Coronary Syndrome (ACS) who are likely to die during their hospitalization. Secondarily, we sought to compare the ability of two procedures to classify subjects into one of two groups (those dying in-hospital and those surviving).

63.2 Materials and Methods

63.2.1 Linear Discriminant Analysis (LDA) and Logistic Regression Analysis (LR)

Linear discriminant analysis (LDA) can be used to determine which variable discriminates between two or more groups of subjects, and to derive a classification model for predicting the group membership of new observations (Tabachnink B.G, Fidell L.S (1996)). In the simplest type of LDA, two-group LDA, a linear discriminant function (LDF) that passes through the centroids of the two groups can be used to discriminate between the two groups.

The LDF is represented by equation:

$$LDF = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

b_j : the value of the j$^{\text{th}}$ coefficient, $j = 1, \ldots, k$.
x_{ij}: the value of the i$^{\text{th}}$ case of the j$^{\text{th}}$ predictor.

The LDF can also be written in standardized form, in which each variable is adjusted by subtraction of its mean value and division by its standard deviation. The standardized coefficients allow you to compare variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability. The value of the standardized form of LDF called discriminant score and it is used to assign objects to groups by using a cut-off for this score. The probability of an event occurring for a given object is calculated as:

$$P(Y_i = 1|X_i) = \frac{1}{1 + (e^{b^T X_i})^{-1}}$$

On the other hand, logistic regression is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome, based on values of a set of predictor variables (Hosmer, D.W. and Lemeshow, S. (1989)). Since the probability of an event must lie between 0 and 1 (for the binary case), it is impractical to model probabilities with linear regression techniques, because the linear regression model allows the dependent variable to take values greater than 1 or less than 0. The logistic regression model is a type of generalized linear model that extends the linear regression model by linking the range of real numbers to the 0-1 range. Define by p_1 the probability of an object is belonging to group 1 and by p_0 the probability of an object belonging to group 0.

The form of the logistic regression model is:

$$z_i = \log(p_{i1}/p_{i0}) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

where
p_{i1}/p_{i0} : is called the odds ratio
b_j : the value of the j$^{\text{th}}$ coefficient, $j = 1, \ldots, k$.
x_{ij} : the value of the i$^{\text{th}}$ case of the j$^{\text{th}}$ predictor

The parameters of the logistic model (b_0 to b_1) are derived by the method of maximum likelihood. From the logistic regression model we can derive the probability of an event occurring as:

$$P(Y_i = 1|X_i) = \frac{e^{b^T X_i}}{1 + (e^{b^T X_i})} = \frac{1}{1 + e^{-b^T X_i}}$$

Using a probability cut-off of 0.5, we can classify an object to group 1 if $p_1 > 0.5$ and to group 0 if $p_1 < 0.5$.

Hence, the two methods do not differ in functional form; they only differ in the estimation of coefficients. Moreover, there are basic differences in the statistical assumptions, which underlie those two methods. With discriminant analysis, the assumptions are: a) the data for the independent variables represent a sample from a multivariate normal distribution. Therefore, predictor variables should be interval or ratio variables, b) The variance of the predictors and the correlations among the predictors within each group should be the same (equal variance/covariance matrices), c) The predictors are not highly correlated with each other. With logistic regression the assumptions are that a logistic regression (i.e. a sigmoidal dependency) exists between the probabilities of group memberships and a linear function of the predictor variables. It is also assumed that observations are independent.

It has been shown that LDA is a more appropriate method when explanatory variables are normally distributed. In the case of categorized variables, LDA remains preferable and fails only when the number of categories is really small (2 or 3). The results of LR, however, are in all these cases constantly close and a little worse than those of LDA. But whenever the assumptions of LDA are not met, the use of LDA is not justified, while LR gives good results since LR can handle both categorical and continues variables, and the predictors do not have to be normally distributed, linearly related or of equal variance within each group regardless of the distribution. (Pohar M, Blas M, Turk S (2004)).

63.2.2 Application in epidemiological data

In this study, we compared the results of discriminant and logistic regression in predicting in-hospital mortality among patients presenting with a range spectrum of acute coronary syndromes. The independent variables which were available as potential predictors for in-hospital mortality was history of coronary heart disease (CHD), hypertension (HTN) and diabetes mellitus (DM), sex, age, body mass index (BMI), smoking habits, initial level of systolic blood pressure (SBP), the estimated creatinine clearance rate (CrCl), the maximum level of MB isoenzyme of creatinine kinase (CPKMBmax) and the time between the onset of symptoms and the admission at the hospital.

Initially, we entered in both discriminant and logistic regression models only the predictors, which were statistically significant in univariate analysis. We used the standardized canonical discriminant function coefficients for discriminant analysis and z statistic (standardized coefficients, Wald statistic) for logistic regression, to evaluate the contribution of each one variable to the discrimination between two groups. The larger the standardized coefficients, the greater are the contribution of the respective variable to the discrimination. We, also, compared the sign and magnitude of coefficients. Secondarily, we performed stepwise discriminant and logistic regression analysis including all available predictors mentioned above. For discriminant analysis, the selection

criterion for entry was the Wilks' Lambda, with a value F-to-enter of 3.84 and a value F-to-remove of 2.71. For logistic regression, was used the set of 0.05 significance levels for entry and 0.1 for removal of variables; these p values were selected to approximate the F values used in the discriminant analysis. We compared the variables selected, the order of selection and the sign and magnitude of coefficients. Equality of the covariance matrices was checked with the Box's M test and it was revealed that they were not equal ($p < 0.001$), thus this assumption for discriminant analysis was not met.

Response operating characteristics (ROC) curves were plotted for each model. An ROC curve graphically displays sensitivity and 100% minus specificity (false positive rate) at several cut-off points. By plotting the ROC curves for two models on the same axes, one is able to determine which test is better for classification, namely, that test whose curve encloses the larger area beneath it. All analyses were performed using the SPSS version 13.0 software.

63.3 Results

Univariate analysis revealed that the CPKMBmax levels, the SBP, the CrCl, gender, age, and DM contribute significantly in the discrimination of patients in those dying during their hospitalization and those surviving. Using in discriminant and logistic regression only these variables; both techniques revealed that CPKMBmax levels, SBP and DM were the most important contributors (Table 1). Moreover, we observe that the direction of the relationships was the same and there were not extreme differences in the magnitude of the coefficients. The correct classification rate was 79% for discriminant analysis and 96.6% for logistic regression. When we used not equal prior probabilities for the two groups the correct classification rate was increased in 96.3%, however in this case decreased the rate of correct classification of patients who died. Figure 1, shows the ROC curves of the aforementioned models, indicating that the logistic model is slightly superior in its classification ability.

The stepwise approach revealed that both models selected the same variables, with the same order of entry (Table 2). Furthermore, the sign of the coefficients were the same and a slight difference was observed in the magnitude of the coefficients. The correct classification rate was 81.4% for discriminant analysis and 96.8% for logistic regression. Figure 2, shows that the logistic model is slightly superior in its classification ability compared to discriminant analysis model.

63.4 Discussion

In general, results from the logistic model agreed with those of discriminant analysis. Both techniques selected the same variables when we performed the

stepwise approach, while entering all significant variables from the univariate analysis in these two methods, only slight differences were observed in the order of predictors (from the most important for the discrimination between the two groups to the less important) between those methods. The overall correct classification rate was good for both, and either would be useful for the prediction of the in-hospital mortality of patients presenting with acute coronary syndromes. Moreover, although the assumption of equal covariance was not hold in this dataset, both methods had similar results.

In conclusion, for this particular problem the logistic regression resulted in the same model, as did discriminant analysis. However, given the slightly better performance of logistic regression, it is preferable to discriminant analysis, particularly when the assumptions are not hold.

Table 1: Variables, standardized and un-standardized coefficients for the discriminant analysis model and logistic regression models.

	Logistic Regression		**Discriminant analysis**	
Predictors	b coefficients	z- statistic	Unstandardized coefficients	Standardized coefficients
CPKMBmax	0.005	4.86	0.007	0.649
SBP	-0.021	3.49	-0.015	-0.390
DM	1.076	3.22	0.812	0.375
CrCl	-0.020	2.55	-0.04	-0.196
Age	0.04	2.14	0.029	0.372
Sex	-0.63	1.87	-0.625	-0.264

DM: history of diabetes mellitus, SBP: initial level of systolic blood pressure, CrCl: the estimated creatinine clearance rate, CPKMBmax: the maximum level of MB isoenzyme of creatinine kinase

Table 2: Variables, standardized and un-standardized coefficients for the discriminant analysis model and logistic regression models, after stepwise approach.

	Logistic Regression		**Discriminant analysis**	
Predictors	b coefficients	z- statistic	Unstandardized coefficients	Standardized coefficients
CPKMBmax	0.005	5.22	0.007	0.692
Age	0.071	3.96	0.036	0.457
SBP	-0.027	3.63	-0.017	-0.411
Sex	-0.938	2.48	-0.805	-0.340
DM	0.901	2.42	0.594	0.274

DM: history of diabetes mellitus, SBP: initial level of systolic blood pressure, CPKMBmax: the maximum level of MB isoenzyme of creatinine kinase

Figure 1: Receiver operating characteristics (ROC) curves for the discriminant analysis model and the logistic regression model.

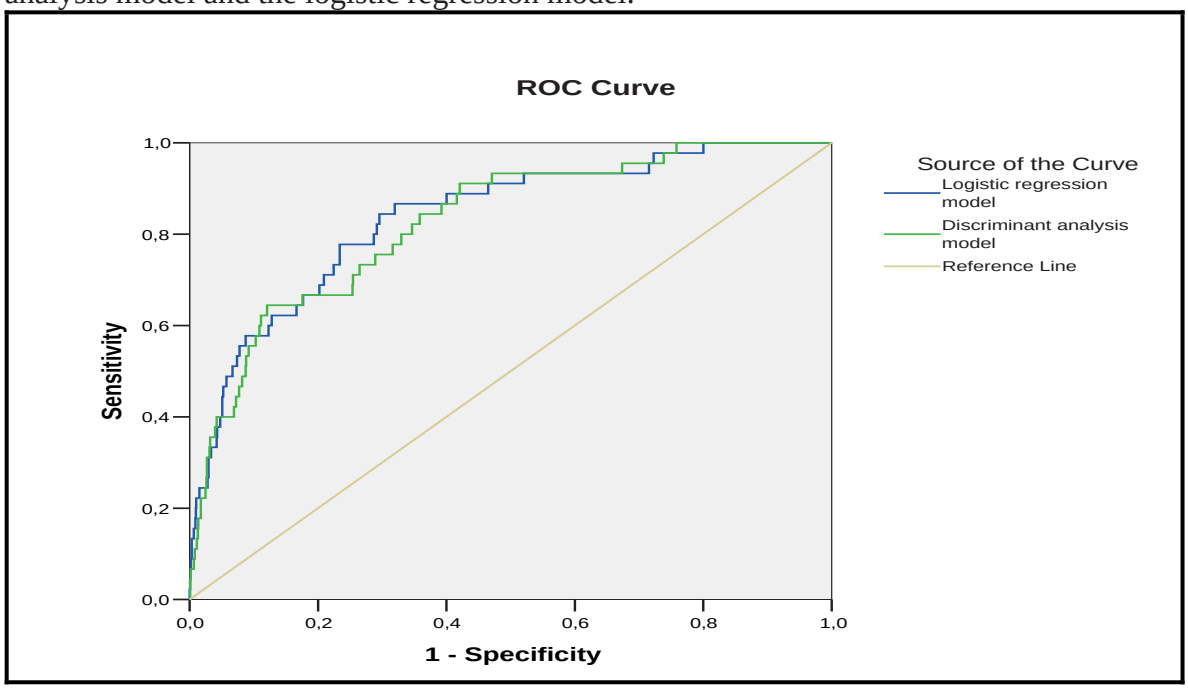
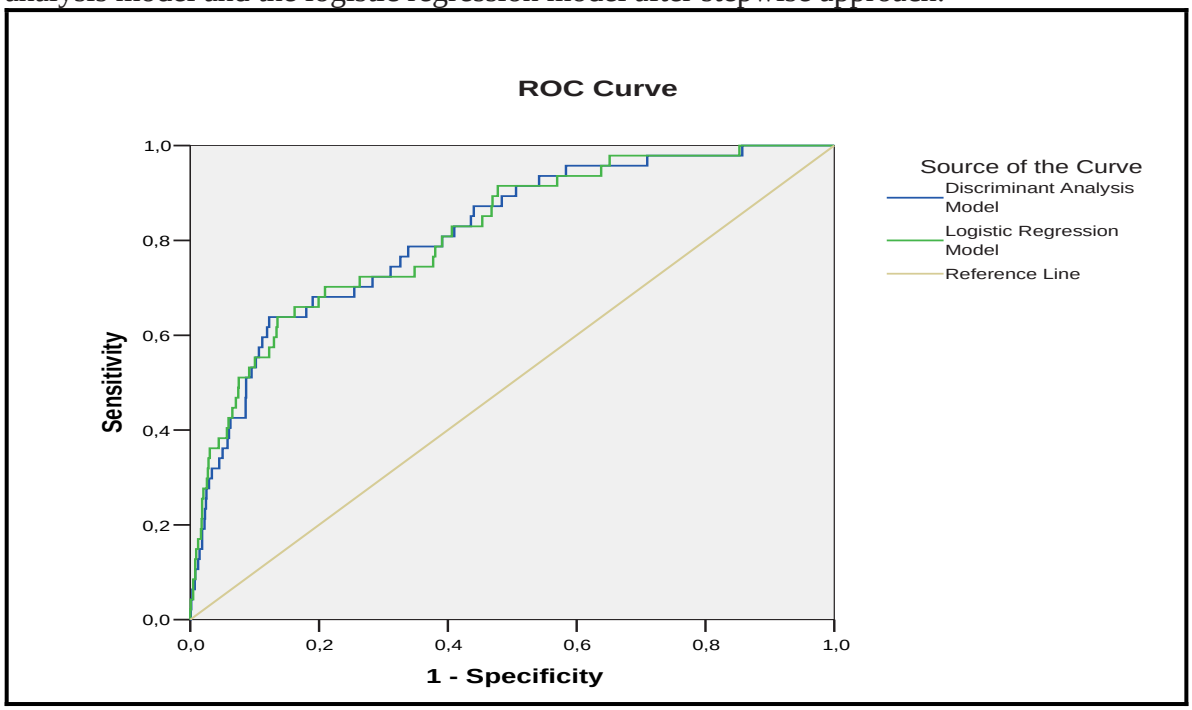


Figure 2: Receiver operating characteristics (ROC) curves for the discriminant analysis model and the logistic regression model after stepwise approach.



References

1. Hosmer, D.W. and Lemeshow, S. (1989): Applied Logistic Regression. New York: Wiley
2. Pohar M, Blas M, Turk S. (2004): Comparison of logistic regression and linear discriminant analysis: A simulation study. Metodoloski Zvezki. 1(1);143-161
3. Tabachnink B.G, Fidell L.S (1996): Using Multivariate Statistics, Harper Collins, New York

64

A Nonparametric Test for an Accelerated Life Time Model

Hannelore Liero

Institute of Mathematics, University of Potsdam, Germany

Abstract: We consider the problem of testing whether survival data can be modelled by an accelerated life time model with a prespecified parametric influence function. Using the relation of an accelerated life time model to a regression model an asymptotic α-test of L_2-type is proposed.

Keywords and phrases: Accelerated life time, survival, nonparametric regression, L_2-type test

64.1 The Problem

We consider a random life time T, which depends on some explanatory variables or covariate X. Such covariate is for example a dose of a drug, temperature or stress. The problem is to draw conclusions about the conditional distribution of T given the covariate, i.e. about the probability that an individual survives the time t when X takes the value x

$$S(t|x) \;=\; \mathsf{P}(T > t|X = x)$$

on the basis of observations $(T_1, X_1), \ldots, (T_n, X_n)$. The function S is the conditional survival function. Note that we assume that the covariates are random.

A popular model for describing the time of survival depending on a covariate is the accelerated life time model, which describes the following situation: Let T_0 be a basic life time whose survival function S_0 does not depend on the covariate x, and let $\psi(\cdot)$ be the function that depicts the influence of the covariates x. If X reduces the life time T by a factor $\psi(X)$ we can write

$$T \;=\; \frac{T_0}{\psi(X)}$$

and the survival function of T is given by

$$S(t|x) = \mathsf{P}(T > t|X = x) = \mathsf{P}(T_0 > t\psi(X)|X = x) = S_0(t\psi(x)). \quad (64.1.1)$$

In this talk we will consider the problem of testing whether S satisfies the accelerated life time model with a prespecified parametric function $\psi(\cdot;\vartheta)$. For this purpose define the following set of accelerated life time survival functions:

$$\mathcal{S} = \{S(\cdot|\cdot) : S(t|x) = S_0(t\psi(x;\vartheta)),\, S_0 \in \mathcal{G}, \vartheta \in \mathbb{R}^d\}$$

where $\mathcal{G}$ is the set of all continuous survival functions and $\psi(\cdot;\vartheta)$ is a (known) function depending on an unknown d-dimensional parameter ϑ. For example, $\psi(\cdot;\vartheta)$ can be a polynomial of degree d. The test problem has the form

$$\mathcal{H}: \quad S \in \mathcal{S} \qquad \text{against} \qquad \mathcal{K}: \quad S \notin \mathcal{S}. \quad (64.1.2)$$

The proposed test procedure is based on the following well-known fact: Set $\mu = \mathsf{E}(\log T_0)$, then a random variable T with survival function (64.1.1) satisfies the equation

$$\log T = -\log(\psi(X)) + \mu + \varepsilon$$

where the random error ε has expectation zero and a variance σ^2 which is independent of X. In other words: The relationship between $Y = \log T$ and X is described by the regression function

$$m(x) = \mathsf{E}(Y|X = x) = -\log(\psi(x)) + \mu.$$

Using this relation the test problem (64.1.2) can be re-formulated into

$$\tilde{\mathcal{H}}: \quad m \in \mathcal{M} \qquad \text{against} \qquad \tilde{\mathcal{K}}: \quad m \notin \mathcal{M}, \quad (64.1.3)$$

where $\mathcal{M} = \{m : m(x) = -\log\psi(x;\vartheta) + \mu,\ \vartheta \in \mathbb{R}^d,\ \mu \in \mathbb{R}\}$ is the parametric class of regression functions.

Thus, we can apply results obtained for testing a regression function in the nonparametric setting to solve the test problem in the underlying survival model.

64.2 The Test Procedure

As test statistic for testing (64.1.3) one has to choose a distance between a good estimator for $m \in \mathcal{M}$ and a good estimator for $m \notin \mathcal{M}$, in other words, a good parametric estimator and a good nonparametric estimator based on smoothing methods.

Let us start with the parametric estimator. It seems to be useful to estimate the unknown parameters in our regression model by the least squares method,

i.e. the estimators $\hat{\mu}$ and $\hat{\vartheta}$ of the parameters of the hypothetical regression are solutions of

$$\min_{\vartheta,\mu} \sum_{i=1}^{n} \left(\log T_i \; - \; \mu + \log \psi(X_i;\vartheta)\right)^2 .$$

Thus, $m \in \mathcal{M}$ is estimated by

$$\hat{m}_n(\cdot, \hat{\vartheta}) = -\log \psi(\cdot;\hat{\vartheta}) + \hat{\mu}.$$

To characterize the regression under the alternative we choose an estimator which is suitable for all possible regression functions, i.e. we apply nonparametric techniques. Nonparametric regression estimators can be written as weighted average of the response variables Y_i, here $Y_i = \log T_i$:

$$\tilde{m}_n(x) \; = \; \sum_{i=1}^{n} W_{b_n i}(x, X_1, \ldots, X_n) Y_i$$

Roughly speaking, the weights $W_{b_n i}$ are chosen such that Y_i gets a large weight, if the corresponding X_i is near the point x, and this, what is "near" is controlled by the smoothing parameter b_n. There are many, many papers on nonparametric estimation of the regression function; as examples we mention Härdle (1990), Fan and Gijbels (1996).

For simplicity of presentation we will suppose that the covariate X is one-dimensional. Furthermore, as nonparametric estimator for m we take the classical Nadaraya-Watson kernel estimator, which is widely investigated. The weights are given by

$$W_{b_n i}(x, X_1, \ldots, X_n) \; = \; \frac{K\left(\frac{x - X_i}{b_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{b_n}\right)}$$

$K : \mathbb{R} \to \mathbb{R}$ is a kernel function, and b_n is a sequence of bandwidths tending to zero as $n \to \infty$.

Under suitable conditions on the distribution of the underlying random variables, on the smoothness of the regression function, on the kernel and the bandwidth several asymptotic properties, such as consistency, asymptotic expression for the mean squared error and limit theorems, are proved. Two of these results are essential for the test procedure considered here: Nonparametric estimators are biased estimators! The distribution of the standardized integrated squared error converges to the standard normal distribution! This leads to the following approach:

- As test statistic the integrated squared distance between the parametric and the nonparametric estimator is proposed.

- Instead of the difference between $\hat{m}_n(\cdot,\hat{\vartheta})$ and $\tilde{m}_n$ we choose the distance between $\tilde{m}_n$ and the smoothed version

$$\tilde{m}_{\mathcal{H}}(x) := \sum_{i=1}^{n} W_{b_n i}(x,\mathbb{X})\hat{m}_n(X_i;\hat{\vartheta}) \qquad \mathbb{X} = (X_1,\ldots,X_n)$$

  of the parametric estimator $\hat{m}_n$ to characterize the hypothesis $\mathcal{H}$. Doing this we avoid problems of the bias.

Thus, a L_2-type test statistic is defined by

$$\begin{aligned} \mathrm{Q}_n &= \int \Big(\tilde{m}_n(x) - \tilde{m}_{\mathcal{H}}(x)\Big)^2 a(x)\,\mathrm{d}x \\ &= \int \Big(\sum_{i=1}^{n} W_{b_n i}(x,\mathbb{X})\big(Y_i - \hat{m}_n(X_i;\hat{\vartheta})\big)\Big)^2 a(x)\,\mathrm{d}x. \end{aligned}$$

Here a is a known weight function, which is introduced to control the region of integration.

To construct the test procedure one has to derive the distribution of Q_n under the null hypothesis. It is impossible to do this exactly. One possibility is to consider the limit distribution of Q_n and to formulate an asymptotic α-test. Based on results of the asymptotic behavior of quadratic forms it was shown that the standardized Q_n is asymptotically normally distributed, that is, under

- smoothness assumptions on the regression function m and the density function of the covariates X_i
- regularity conditions on the kernel K and the bandwidth b_n
- conditions ensuring the consistency of the l.s.e. for ϑ

we have

$$n b_n^{1/2}\,(\mathrm{Q}_n - e_n) \to \mathsf{N}(0,\tau^2). \tag{64.2.4}$$

Here

$$e_n = (nb_n)^{-1}\sigma^2 \int g^{-1}(x)\,a(x)\,\mathrm{d}x\,\kappa_1 \quad \text{and} \quad \tau^2 = 2\sigma^2 \int g^{-2}(x)a^2(x)\,\mathrm{d}x\,\kappa_2,$$

where $\kappa_1 = \int K^2(x)\,\mathrm{d}x$ and $\kappa_2 = \int (K*K)^2(x)\,\mathrm{d}x$ are constants depending on the kernel K and g denotes the density of X_i. This result can be found in several papers and books about nonparametric regression estimation. A detailed formulation is given in Liero (1992), see also Härdle and Mammen (1993).

Now, let us consider the test problem: To apply (64.2.4) for the construction of the test we have to replace the unknown term σ^2 and the marginal density

g in e_n and τ^2 by suitable estimators. For the estimation of g one should use a nonparametric kernel estimate, for the estimation of σ^2 one can apply a method for estimating the variance in a homoscedastic nonparametric regression model with random design considered in Liero (2003).

Suppose, we have chosen good estimators $\hat{\sigma}^2$ and $\hat{g}_n$ for these unknown terms, then an asymptotic α- test is given by the rule:

Reject the null hypothesis $m \in \mathcal{M}$ respectively $S \in \mathcal{S}$ if

$$\mathrm{Q}_n \geq (nb_n^{1/2})^{-1}\, \hat{\tau}_n\, z_\alpha \;+\; \hat{e}_n,$$

where

$$\hat{e}_n = (nb_n)^{-1}\hat{\sigma}_n^2 \int \hat{g}_n^{-1}(x)\, a(x)\, \mathrm{d}x\, \kappa_1, \qquad \hat{\tau}_n^2 = 2\,\hat{\sigma}_n^2 \int \hat{g}_n^{-2}(x) a^2(x)\, \mathrm{d}x\, \kappa_2,$$

and z_α is the $(1-\alpha)$-quantile of $\mathsf{N}(0,1)$.

64.3 Closing Remarks and Open Problems

The procedure given above is a proposal for testing whether the underlying data can be modelled by an accelerated life time model. To justify this approach one has to study properties of this test. One important point is the power under local alternatives. The asymptotic behavior of the power under local alternatives in the regression model is investigated. These results have to be translated into results for the power under alternatives defined in the survival model.

Secondly, here we considered the uncensored case. Very often in survival analysis we have censored data. Is there an appropriate modification of the procedure for this case?

And finally, it is not clear, whether the approximation of the distribution of Q_n by its limit distribution is sufficiently good. So, the question arises, whether resampling methods are a good alternative for the determination of critical values for the test in practical situations.

References

1. Fan, J. and Gijbel, I. (1996). *Local Polynomial Modelling and Its Applications* Chapman and Hall

2. Härdle, W. (1990) *Applied Nonparametric Regression* Cambridge University Press, Cambridge

3. Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits, *Annals of Statistics*, **21**, 1926–1947

4. Liero, H. (1992). Asymptotic normality of a weighted integrated square error of kernel regression estimates with data-dependent bandwidth, *J. Statistical Planning and Inference* **30**, 307-325

5. Liero, H. (2003). Testing homoscedasticity in nonparametric regression, *Nonparametric Statistics*, **15**, 31-51

65

Markov Reward Model for Multi-State System Reliability Assessment

Anatoly Lisniaski[1,2], Ilia Frenkel[2], Lev Khvatskin[2] and Yi Ding[2]

[1] *Israel Electric Corporation Ltd., Israel*
[2] *International Reliability and Risk Management Center (IRRMC), Sami Shamoon College of Engineering, Israel*

Abstract: The paper considers reliability measures for multi-state system where the system and its components can have different performance levels ranged from perfect functioning up to complete failure. The suggested approach presents a generalized reliability measure as a functional of trajectories of two stochastic processes - output performance of the entire multi-state system and corresponding demand. It is shown how the commonly used reliability measures can be derived from this functional. The procedure for the system reliability measures computation is based on the Markov Reward Model. Numerical example is presented in order to illustrate the approach.

Keywords and phrases: Multi-state system, reliability measure, Markov reward model

65.1 Introduction

Traditional binary-state reliability models allow for a system and its components only two possible states: perfect functionality (Up) and complete failure (Down). However, many real-world systems are composed of multi-state components, which have different performance levels and for which one cannot formulate an "all or nothing" type of failure criterion. Failures of some system elements lead in these cases only to the performance degradation. Such systems are called Multi-state Systems (MSS). The traditional reliability theory, which is based on a binary approach, has recently been extended by allowing components and system to have an arbitrary finite number of states. According to generic Multi-state System (MSS) model (Lisnianski and Levitin (2003)), any system element $j \in \{1, 2, \ldots, n\}$ can have k_j different states corresponding to the performance rates, represented by the set $\mathbf{g_j} = \left\{g_{j1}, g_{j2}, \ldots, g_{jk_j}\right\}$, where g_{ji} is the performance rate of element j in the state i, $i \in \{1, 2, \ldots, k\}$.

The performance rate $G_j(t)$ of element j at any instant $t \geq 0$ is a discrete-state continuous-time stochastic process that takes its values from $\mathbf{g_j} : G(t) \in \mathbf{g_j}$. The system structure function $G(t) = \phi(G_1(t), \ldots, G_n(t))$ produces the stochastic process corresponding to the output performance of the entire MSS. In practice a desired level of system performance (demand) also can be represented by a discrete-state continuous-time stochastic process $W(t)$. The relation between the MSS output performance and the demand represented by two corresponding stochastic processes should be studied in order to define reliability measures for the entire MSS.

The list of MSS reliability measures, that were introduced till now, one can find in Aven (1993). In practice the most commonly used MSS reliability measures are probability of failure-free operation during time interval $[0, t]$ or MSS Reliability Function $R(t)$, MSS instantaneous (point) availability, mean time to MSS failure, mean accumulated performance deficiency for a fixed time interval $[0, t]$, etc.

In the paper generalized approach for the computation of main MSS reliability measures was suggested. The approach is based on application of the Markov Reward Model. The main MSS reliability measures can be found by corresponding rewards definitions for this model and then by using standard procedure for finding expected accumulated reward during time interval $[0, t]$ as a solution of system of differential equations.

65.2 Model description

65.2.1 Generalized MSS Reliability Measure

The MSS behavior is characterized by its evolution in the space of states. The entire set of possible system states can be divided into two disjoint subsets corresponding to acceptable and unacceptable system functioning. MSS entrance into the subset of unacceptable states constitutes a failure. The system state acceptability depends on the relation between the MSS output performance and the desired level of this performance - demand $W(t)$ - that is determined outside of the $W(t)$ is also a random process that can take discrete values from the set $w = \{w_1, \ldots, w_M\}$. The desired relation between the system performance and the demand at any time instant t can be expressed by the acceptability function $\Phi(G(t), W(t))$. In many practical cases, the MSS performance should be equal or exceed the demand. So, in such cases the acceptability function takes the following form

$$\Phi(G(t), W(t)) = G(t) - W(t) \tag{1}$$

and the criterion of state acceptability can be expressed as $\Phi(G(t), W(t)) \geq 0$. A general expression defining MSS reliability measures can be written in the

following form

$$R = E\{F[\Phi(G(t), W(t))]\} \tag{2}$$

where E - expectation symbol, F - functional that determines corresponding type of reliability measure, Φ - acceptability function.

Many important MSS reliability measures can be derived from the expression (2) depending on functional F that may be determined by different ways. It may be a probability $\Pr\{\Phi(G(t), W(t)) \geq 0\}$ that within specified time interval $[0, t]$ the acceptability function (1) will be nonnegative. This probability characterizes MSS availability. It may be also a time up to MSS first entrance into the set of unacceptable states, where $\Phi(G(t), W(t)) < 0$, a number of such entrances within time interval $[0, t]$ etc. If the acceptability function is defined as $F(\Phi(G(t), W(t)) = W(t) - G(t)$, if $W(t) > G(t)$ and $F(\Phi(G(t), W(t)) = 0$ if $W(t) \leq G(t)$ a functional $F(\Phi(G(t), W(t)) = \int\limits_0^t \Phi(G(t), W(t))dt$ will characterize an accumulated performance deficiency during time interval $[0, t]$. In the paper generalized approach for main reliability measures computation is suggested.

65.2.2 Markov Reward Model: General Description

General Markov reward model considers the continuous time Markov chain with set of states $\{1, \ldots, k\}$ and transition intensity matrix $\mathbf{a} = |a_{ij}|, i, j = 1, \ldots, k$. It is assumed the while the process is in any state i during any time unit some money r_{ii} should be paid. It is also assumed that if there is a transition from state i to state j the amount r_{ij} will be paid. The amounts r_{ii} and r_{ij} are called rewards. They can be negative while representing loss or penalty. The main problem is to fined total expected reward accumulated up to time instant T under specific initial conditions. Let $V_i(t)$ be the total expected reward accumulated up to time t at state i. According to Howard (1960), the following system of differential equations must be solved under initial conditions $V_i(0) = 0,\ i = 1, \ldots, k$ in order to find the total expected reward:

$$\frac{dV_i(t)}{dt} = r_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{k} a_{ij} r_{ij} + \sum_{j=1}^{k} a_{ij} V_j(t), i = 1, ..., k \tag{3}$$

65.2.3 Rewards Determination for MSS Reliability Computation

MSS instantaneous (point) availability $A(t)$ is the probability that the MSS at instant $t > 0$ is in one of the acceptable states:

$$A(t) = \Pr\{\Phi(G(t), W(t)) \geq 0\}.$$

$A(t)$ is defined as mean part of time, when the system is staying in the set of acceptable states during time interval $[0, t]$. In order to assess $A(t)$ for MSS the rewards in matrix $\mathbf{r}$ for MSS model should be determined by the following manner.

- The rewards associated with all acceptable states should be defined as 1.
- The rewards associated with all unacceptable states should be zeroed as well as all rewards associated with transitions.

The mean reward $V_K(t)$ accumulated during interval $[0, t]$ will define a part of time that MSS will be in the set of acceptable states in the case when the state K is the initial state. This reward should be found as a solution of system (3). After solving the (3) and finding $V_K(t)$, MSS instantaneous availability can be obtained as $A(t) = V_K(t)/t$.

Mean number $N_f(t)$ of MSS failures during time interval $[0, t]$. This measure can be treated as mean number of MSS entrances the set of unacceptable states during time interval $[0, t]$. For its computation the rewards associated with each transition from the set of acceptable states to the set of unacceptable states should be defined as 1. All other rewards should be zeroed. In this case mean accumulated reward $V_K(t)$ will define mean number of entrances in unacceptable area during time interval $[0, t] : N_f(t) = V_K(t)$.

Mean Time To Failure (MTTF) is the mean time up to the instant when the MSS enters the subset of unacceptable states for the first time. For its computation the combined performance-demand model should be transformed - all transitions that return MSS from unacceptable states should be forbidden, because for this case all unacceptable states should be treated as absorbing states.

In order to assess MTTF for MSS the rewards in matrix $\mathbf{r}$ for the transformed performance-demand model should be determined by the following manner.

- The rewards associated with all acceptable states should be defined as 1.
- The reward associated with unacceptable (absorbing) states should be zeroed as well as all rewards associated with transitions.

In this case mean accumulated reward $V_K(t)$ will define mean time accumulated up to the first entrance into the subset of unacceptable states or MTTF.

Probability of MSS failure during time interval $[0, t]$. Model should be transformed as in the previous case - all unacceptable states should be treated as absorbing states and, therefore, all transitions that return MSS from unacceptable states should be forbidden. Rewards associated with all transitions to the absorbing state should be defined as 1. All other rewards should be zeroed. Mean accumulated reward $V_K(t)$ will define for this case probability of MSS

failure during time interval $[0, t]$. Therefore, MSS reliability function can be obtained as $R(t) = 1 - V_K(t)$.

65.3 Numerical Example

Consider the Air conditioning system, used in hospital. The system consists of two 5 years old main conditioners and one conditioner in cold reserve. The reserve conditioner begins to work only when one of the main conditioners has failed. Conditioners failure and repair rates: $\lambda = \lambda^\star = 10, \mu = \mu^\star = 100$. The state-space diagram for this system is presented in Fig. 1.

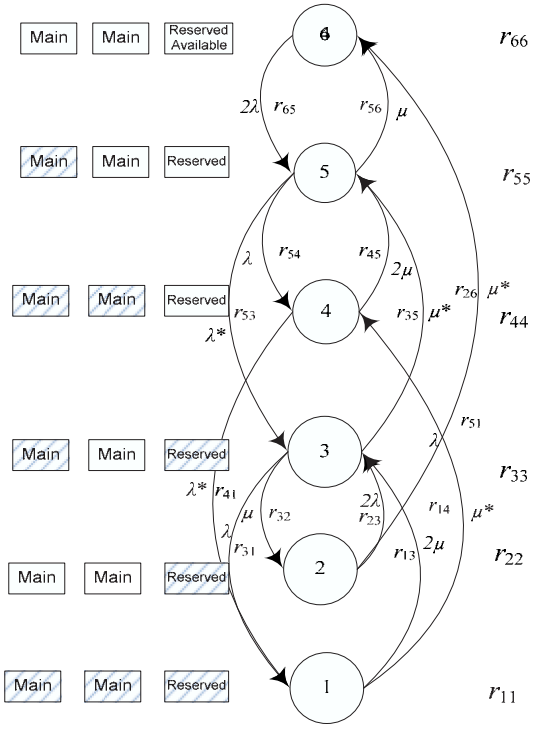


Fig. 1 System with two on-line conditioners and one conditioner in cold reserve

State 6 - The 2 main conditioners are on-line and the reserved conditioner is available. The system performance $g_6 = 2$. State 5 - One of the main conditioners is failed and replaced by reserved conditioner. The system performance $g_5 = 2$. State 4 - The second main conditioner is failed, only reserved conditioner is on-line. The system performance $g_4 = 1$. State 3 - The reserved conditioner is failed, only one main conditioner is on-line. The system performance $g_3 = 1$. State 2 - The reserved conditioner is failed, two main conditioners are on-line. The system performance $g_2 = 2$. State 1 - Full system failure. The system performance $g_1 = 0$.

We shall find MSS reliability measures for constant demand level $w = 1$. Under this condition there is only one unacceptable state - state 1. The transition intensity matrix is as follows.

$$\alpha = \begin{array}{|cccccc|} -(2\mu+\mu^\star) & 0 & 2\mu & \mu^\star & 0 & 0 \\ 0 & -(2\lambda+\mu^\star) & 2\lambda & 0 & 0 & \mu^\star \\ \lambda & \mu & -(\lambda+\mu+\mu^\star) & 0 & \mu^\star & 0 \\ \lambda^\star & 0 & 0 & -(\lambda^\star+2\mu) & 2\mu & 0 \\ 0 & 0 & \lambda^\star & \lambda & -(\lambda+\lambda^\star+\mu) & \mu \\ 0 & 0 & 0 & 0 & 2\lambda & -2\lambda \end{array} \quad (4)$$

In order to find the MSS instantaneous (point) availability $A(t)$ and the mean total number of system failures $N_f(t)$ we should present the reward matrixes r_A and r_N accordingly in the following form

$$r_A = |r_{ij}| = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} \qquad r_N = |r_{ij}| = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix} \quad (5)$$

By solving the system of differential equations (3) with transition intensity matrix (4) and reward matrices (5) we can obtain MSS point availability and mean total number of system failures. The results of calculation are presented in Fig.2 and Fig. 3.

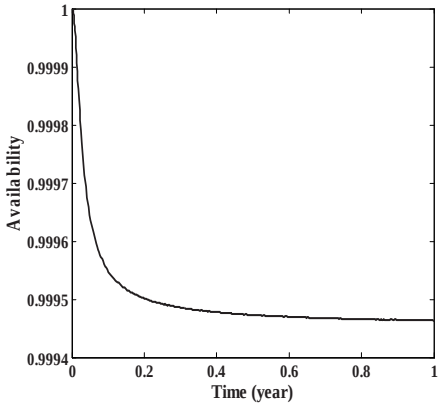


Fig. 2 Calculation the MSS instantaneous (point) availability

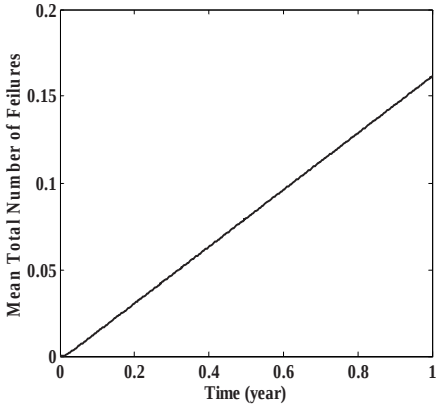


Fig. 3 Mean Total Number of System Failures

65.4 Conclusions

1. Generalized reliability measure for MSS that is an expectation of functional from two stochastic processes - MSS output performance G(t) and demand W(t) - was suggested in the paper. It was shown that many of usually used

in practice MSS reliability measures can be easily derived from this generalized measure.
2. The universal method was suggested to compute main MSS reliability measures. The method is based on different reward matrix determination for MSS model that is interpreted as Markov reward model.
3. The approach suggested is well formalized and suitable for practical application in reliability engineering. 4. The numerical example is presented in order to illustrate the suggested approach.

References

1. Aven, T. (1993). On performance measures for multi-state monotone systems, *Reliability Engineering and System Safety*, **41**, 259-266.
2. Aven, T. and Jensen, U. (1999). *Stochastic Models in Reliability*, Springer, NY.
3. Carrasco, J. (2003). Markovian Dependability/Performability Modeling of Fault-tolerant Systems, in *Handbook of Reliability Engineering* (Ed. H. Pham) Springer, London, Berlin, NY.
4. Howard, R. (1960). *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, Massachusetts.
5. Lisnianski, A. and Levitin, G. (2003). *Multi-state System Reliability. Assessment, Optimization and Applications*, World Scientific, NJ, London, Singapore.

66

Unbiased Estimators for the Multivariate Pólya and Wishart Distributions

Ya. Lumelskii[1], V. Voinov[2], M. Nikulin[3] and P. Feigin[1]

[1]*Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel*
[2]*Kazakhstan Institute of Management, Economics and Strategic Research, Almaty, Kazakhstan*
[3]*EA 2961, Statistique Mathematique Universite Bordeaux 2, Bordeaux, France, V.A. Steklov Mathematical Institute, St.Petersburg, Russia*

Abstract: A generalization of the classical random sampling scheme is suggested. Unbiased estimators for functions of unknown parameters for the multivariate Pólya, and the Wishart distributions are derived.

Keywords and phrases: Unbiased estimators, multivariate Pólya distribution, Wishart distribution

66.1 Introduction

Problems of unbiased estimation are typically solved by using a sample of n independent identically distributed (i.i.d.) random variables or vectors (see, e.g., Johnson *et al.* (1997), Kotz *et al.* (2003), Nikulin and Voinov (1996), Voinov and Nikulin (1996)). This sampling strategy may be called the classical random sampling. Unfortunately, this scheme is inapplicable for many discrete probability distributions. In particular, it is inapplicable for the multivariate Pólya distribution, since for n independent Pólya distributed vectors their sum will not follow the Pólya distribution.

If a family of distributions possesses a complete sufficient statistic, the classical random sampling scheme can be generalized. In this talk a generalization of the classical random sampling scheme is suggested. In some sense, this generalization is related to random walks in the space of a sufficient statistic (see, e.g., Lumelskii (1973)), Lumelskii (1998)) however, the method can be introduced without bringing the concept of random walks altogether. We consider some applications of the above concept to the multivariate Pólya, and the Wishart distributions.

Basing on the proposed generalization many new minimum variance unbiased estimators for functions of unknown parameters for the multivariate Pólya and

the Wishart distributions are constructed.

66.2 Multivariate Pólya distribution

The multivariate Pólya distribution is an important probability models, which has numerous applications (see, Janardan and Patil (1972), Johnson *et al.* (1997), Lumelskii (1973), and others). For the bibliography and applications of this distribution see the monograph Johnson *et al.* (1997).
The multivariate Pólya distribution with parameters m, $\mathbf{p}$ and λ is defined as

$$P(\mathbf{x}; m, \mathbf{p}, \lambda) = \binom{m}{x_1, ..., x_k} \frac{\prod_{i=1}^{k} p_i^{[x_i;\lambda]}}{1^{[m;\lambda]}}, \tag{66.2.1}$$

where

$$a^{[r;\lambda]} = \prod_{h=0}^{r-1}(a + \lambda h);\ a^{[0;\lambda]} = 1,$$

$$\mathbf{p} = (p_1, p_2, \ldots, p_k)^{'}, p_k = 1 - \sum_{i=1}^{k-1} p_i, 0 < p_i < 1,$$

$\mathbf{x}$ is a random vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)^{'}$ such that $x_k = m - \sum_{i=1}^{k-1} x_i$, λ is a real valued constant ($p_i + \lambda(m-1) > 0$ for all i) , m and x_i being nonnegative integers.
Suppose a random vector $\mathbf{U} = (U_1, ..., U_k)^{'}$, $U_k = M - \sum_{i=1}^{k-1} U_i$, based on $M > m$ observations from the population distribution (66.2.1) is given. Parameters m, M and λ are considered to be known. The vector $\mathbf{U}$ obviously possesses the multivariate Pólya sampling distribution

$$P(\mathbf{U}; M, \mathbf{p}, \lambda) = \binom{M}{U_1, ..., U_k} \frac{\prod_{i=1}^{k} p_i^{[U_i;\lambda]}}{1^{[M;\lambda]}}. \tag{66.2.2}$$

Thus the distribution of $\mathbf{U}$ under the discussed sampling scheme coincides with the population distribution (66.2.1) with parameter m replaced by M, the total number of observations from (66.2.1). Since the multiplier $\prod_{i=1}^{k} p_i^{[U_i;\lambda]}$ depends on unknown parameters $p_i,\ i = 1, ..., k$, through the vector $\mathbf{U}$, from the definition of a sufficient statistic it follows that the random vector $\mathbf{U}$ is the sufficient for p_i statistic. Assuming that the parametric space $\Delta(\mathbf{p})$ of the family (66.2.2) contains a $k-1$-dimensional parallelepiped the following is valid.

Proposition 66.2.1 *The UMVUE* $\hat{P}(\mathbf{x}; m, \mathbf{p}, \lambda)$ *of the probability (66.2.1) is*

$$\hat{P}(\mathbf{x}; m, \mathbf{p}, \lambda) = \frac{\prod_{i=1}^{k} \binom{U_i}{x_i}}{\binom{M}{m}}. \tag{66.2.3}$$

If $\lambda = 0$, then (66.2.1) represents the multinomial probability distribution

$$P(\mathbf{x}; m, \mathbf{p}) = \binom{m}{x_1, ..., x_k} \prod_{i=1}^{k} p_i^{x_i}. \tag{66.2.4}$$

If

$$p_i = \frac{\theta_i}{N},\ 0 \le \theta_i < N,\ \sum_{i=1}^{k} \theta_i = N,\ \lambda = -\frac{1}{N},$$

then (66.2.1) reduces to the multivariate hypergeometric probability distribution

$$P_\theta(\mathbf{x}; m) = \frac{\prod_{i=1}^{k} \binom{\theta_i}{x_i}}{\binom{N}{m}},$$

The UMVUE (66.2.3) does not depend on parameter λ, which implies the following.

Corollary 66.2.1 *Let the random vector* $\mathbf{U} = (U_1, ..., U_k)^{'}$ *have the multinomial distribution* $P(\mathbf{U}; M, \mathbf{p})$ *or the multivariate hypergeometric distribution* $P_\theta(\mathbf{U}; M)$*, then their UMVUEs are defined by the same formula (66.2.3).*

Note that (see, Nikulin and Voinov (1996), Voinov and Nikulin (1996))) for the multinomial distribution the classical random sampling scheme is applicable: $M = n \times m,\ U = \sum_{j=1}^{n} X_j;\ U_i = \sum_{j=1}^{n} X_{ij}$. U has a multinomial distribution $P(\mathbf{U}; n \times m, \mathbf{p})$. Using (66.2.3), we obtain the UMVUE in the form

$$\hat{P}(\mathbf{x}; m, \mathbf{p}) = \prod_{i=1}^{k} \binom{U_i}{x_i} \left[\binom{nm}{m}\right]^{-1}. \tag{66.2.5}$$

Example 66.2.1 *Let independent random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *have multivariate Pólya distributions (66.2.1) with parameters* m_x, $\mathbf{p}_x$, λ_x *and* m_y, $\mathbf{p}_y$, λ_y, *dimensions of vectors* $\mathbf{X}$, $\mathbf{a}$ *and* $\mathbf{Y}$, $\mathbf{b}$ *being* k_x *and* k_y, *respectively. Given statistics* $\mathbf{U}_x$, $\mathbf{U}_y$, $M_x > m_x$, $M_y > m_y$, *we can construct the UMVUE* $\hat{P}(\xi_1 > 0)$ *of the probability*

$$P(\mathbf{a}^{'}\mathbf{X}+\mathbf{b}^{'}\mathbf{Y}+c > 0) \equiv P(\xi_1 > 0) = \sum_{\mathbf{a}'\mathbf{x}+\mathbf{b}'\mathbf{y}+c>0} P(\mathbf{x}; m_x, \mathbf{p}_x, \lambda_x) P(\mathbf{y}; m_y, \mathbf{p}_y, \lambda_y).$$

as

$$\hat{P}(\xi_1 > 0) = \sum_{\mathbf{a}'\mathbf{x}+\mathbf{b}'\mathbf{y}+c>0} \hat{P}(\mathbf{x}; m_x, \mathbf{p}_x, \lambda_x)\ \hat{P}(\mathbf{y}; m_y, \mathbf{p}_y, \lambda_y), \tag{66.2.6}$$

where estimators $\hat{P}(\mathbf{x}; m_x, \mathbf{p}_x, \lambda_x)$ *and* $\hat{P}(\mathbf{y}; m_y, \mathbf{p}_y, \lambda_y)$ *are defined by formula (66.2.3).*

If $k_x = k_y = 1$, $\lambda_x = \lambda_y = 0$ *and* $c = 0$, $a = 1$, $b = -1$, *then the probability* $R \equiv P(\xi_1 > 0) = P(X < Y)$ *has the following form:*

$$R = \sum_{x=0}^{m} \sum_{y=x+1}^{m_y} \binom{m_x}{x} {p_x}^x (1-p_x)^{m_x - x} \binom{m_y}{y} {p_y}^y (1-p_y)^{m_y - y} h(m_y - x - 1),$$

where $m = \min(m_x, m_y)$ *and* h *is the Heaviside function* $h(z) = 1$, *if* $z \geq 0$ *and* $h(z) = 0$, *if* $z < 0$.
The UMVUE for R *is defined by the formula:*

$$\hat{R} = \sum_{x=0}^{m} \sum_{y=x+1}^{m_y} \frac{\binom{m_x}{x}\binom{M_x - m_x}{U_x - x}}{\binom{M_x}{U_x}} \frac{\binom{m_y}{y}\binom{M_y - m_y}{U_y - y}}{\binom{M_y}{U_y}} h(m_y - x - 1).$$

Example 66.2.2 *Let* $\mathbf{U}$ *be a random vector, which has the multivariate Pólya distribution* $P(\mathbf{U}; M, \mathbf{p}, \lambda)$. *Let us construct unbiased estimators of the probability* p_i *and the variance* $Var(\hat{p}_i)$, $i = 1, ..., k$, *where* M *and* λ *are known, and* $M > 2$. *Assuming in (66.2.1)* $m = 1$ *and* $\mathbf{x}$ *be such a vector, that its* i*-th component is equal to one, while other components vanish, we have* $P(\mathbf{x}; 1, \mathbf{p}, \lambda) = p_i$. *According to the formula (66.2.3), the unbiased estimator of* p_i *is*

$$\overset{\wedge}{p_i} = \frac{U_i}{M}. \tag{66.2.7}$$

Similarly, we have

$$P(\mathbf{x}; 2, \mathbf{p}, \lambda) = \frac{p_i(p_i + \lambda)}{1 + \lambda} = \frac{1}{1 + \lambda} \left[{p_i}^2 + p_i \lambda \right].$$

By using (66.2.3) and (66.2.7) we obtain the unbiased estimator for ${p_i}^2$ *as follows:*

$$\overset{\wedge}{{p_i}^2} = (1 + \lambda) \frac{U_i(U_i - 1)}{M(M-1)} - \lambda \frac{U_i}{M}. \tag{66.2.8}$$

Applying (66.2.7) and (66.2.8) we obtain the estimator $\hat{V}ar(\hat{p}_i)$ *of* $Var(\hat{p}_i)$ *as*

$$\hat{V}ar(\hat{p}_i) = (\hat{p}_i)^2 - \overset{\wedge}{{p_i}^2} = \frac{U_i(M - U_i)(1 + M\lambda)}{M^2(M-1)}. \tag{66.2.9}$$

Unlike (66.2.3), the estimator (66.2.9) depends on parameter λ.

66.3 The Wishart probability distribution

The notation

$$\mathbf{X} \sim W(\mathbf{\Sigma}, k, r), \quad r > k,$$

means that a random $k \times k$ positive definite matrix $\mathbf{X}$ has the Wishart distribution with covariance matrix $\mathbf{\Sigma}$ and r degrees of freedom (see, Leung and Chan (1998) and others).
Its density function is

$$w(\mathbf{x}; \mathbf{\Sigma}, k, r) = C(k, r)[\det \mathbf{\Sigma}]^{-\frac{r}{2}}[\det \mathbf{x}]^{\frac{r-k-1}{2}} \exp\{-0.5\mathrm{tr}(\mathbf{x}\mathbf{\Sigma}^{-1})\}, \quad (66.3.10)$$

where

$$C(k, r) = \left[2^{\frac{rk}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{j=1}^{k} \Gamma\left(\frac{r+1-j}{2}\right)\right]^{-1}.$$

Everywhere below we assume that the properties of the parametric space imply completeness of the family of the Wishart distributions.

Proposition 66.3.1 *Let positive definite random matrix* $\mathbf{Y}$ *have the Wishart distribution* $W(\mathbf{\Sigma}; k, N)$ *and* $N > r + k + 1$. *Then the UMVUE of density function (66.3.10) has the following form:*

$$\hat{w}(\mathbf{x}; \mathbf{\Sigma}, k, r) = \frac{C(k, r)C(k, N-r)\,[\det \mathbf{x}]^{\frac{r-k-1}{2}}\,[\det(\mathbf{Y}-\mathbf{x})]^{\frac{N-r-k-1}{2}}}{C(k, N)\,[\det \mathbf{Y}]^{\frac{N-k-1}{2}}}, \quad (66.3.11)$$

if matrices $\mathbf{x}, \mathbf{Y}, \mathbf{Y}-\mathbf{x}$ *are positive definite (p.d.), and zero otherwise.*

Corollary 66.3.1 *Let* $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ *be i.i.d sample* $k \times k$ *random matrices,* $\mathbf{X}_i \sim W(\mathbf{\Sigma}, k, r)$, $i = 1, ..., n$. *Under this classical random sampling scheme the UMVUE (66.3.11) becomes*

$$\hat{w}(\mathbf{x}; \mathbf{\Sigma}, k, r) = \frac{C(k, r)C(k, (n-1)r)\,[\det \mathbf{x}]^{\frac{r-k-1}{2}}\,[\det(\mathbf{Y}_n-\mathbf{x})]^{\frac{(n-1)r-k-1}{2}}}{C(k, nr)\,[\det \mathbf{Y}_n]^{\frac{nr-k-1}{2}}}, \quad (66.3.12)$$

where the sufficient statistic

$$\mathbf{Y}_n = \sum_{i=1}^{n} \mathbf{X}_i \sim \mathbf{W}(\mathbf{\Sigma}, k, nr).$$

Consider the problem of unbiased estimating of the moment generating function for the Wishart probability distribution. To the best of our knowledge it is also an open problem in the case of the classical random sampling scheme.

Proposition 66.3.2 *The UMVUE* $\hat{H}(\mathbf{\Omega})$ *of the moment generating function*

$$H(\mathbf{\Omega}) = E_{\mathbf{\Sigma}} \exp \mathrm{tr}(\mathbf{\Omega}\mathbf{X}) = \int_{\mathcal{A}} \exp\{\mathrm{tr}(\mathbf{\Omega}\mathbf{x})\} w(\mathbf{x}; \mathbf{\Sigma}, k, r) d\mathbf{x}, \quad (66.3.13)$$

of the Wishart distribution is

$$\hat{H}(\mathbf{\Omega}) = {}_1F_1\left(\frac{r}{2}; \frac{N}{2}; -\mathbf{UYU}'\right), \tag{66.3.14}$$

where ${}_1F_1$ *is the confluent hypergeometric function of the matrix argument,* $\mathbf{\Omega}_1 = \mathbf{U}'\mathbf{U} = -\mathbf{\Omega}$ *is a positive definite matrix, and* $\mathbf{Y}$ *is a sufficient statistic for parameters of (66.3.10).*

References

1. Janardan, K.G. and Patil, G.P. (1972). A unified approach for a class of multivariare hypergeometric models, *Sankhya A*, **34**, 363–376.

2. Johnson, N.L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, John Wiley and Sons.

3. Kotz, S., Lumelskii, Ya. and Pensky, M. (2003). *The stress-strength model and its generalizations. Theory and applications*, World Scientific Publishing.

4. Leung, P.L. and Chan, W.Y. (1998). Estimation of the scale matrix and its eigenvalues in the Wishart and the multivariate F distributions, *Ann. Inst. Statist. Math.*, **50**, 523–530.

5. Lumelskii, Ya.P. (1973). Random walks related to generalized urn schemes, *Soviet Math. Dokl.*, **14**, 628–632.

6. Lumelskii, Ya.P. (1998). Random Walks in the Space of a Sufficient Statistic and Statistical Inference, *Journ. Math. Sciences*, **88**, 862–870.

7. Nikulin, M.S. and Voinov, V.G. (1996). Tables of the best possible unbiased estimates for functions of parameters of multinomial and negative multinomial distributions, *Journ. Math. Sciences*, **81**, 2363–2367.

8. Voinov, V.G. and Nikulin, M.S. (1996). *Unbiased Estimators and their Applications. Volume 2: Multivariate case.*, Kluwer Academic Publishers.

67

Fitting Frailty Models via Linear Mixed Models Using Model Transformation

Goele Massonnet, Paul Janssen and Tomasz Burzykowski

Hasselt University, Center for Statistics, Agoralaan 1, B-3590 Diepenbeek, Belgium

Abstract: Frailty models are widely used to model clustered survival data. Classical ways to fit frailty models are likelihood based. We propose an alternative approach in which the original problem of 'fitting a frailty model' is reformulated into the problem of 'fitting a mixed model' using model transformation. Based on a simulation study, we show that the proposed method provides a good and simple alternative for fitting frailty models for data sets with a sufficiently large number of clusters and moderate to large sample sizes within clusters.

Keywords and phrases: Frailty model, random treatment by center interaction, model transformation, linear mixed model

67.1 Introduction

Frailty models are widely used to fit clustered survival data. Data from multicenter clinical trials are a typical example of clustered data; data within the same center all share the same random cluster effect. The shared frailty model provides an appropriate way to describe the within cluster dependence of outcomes. Classical ways to fit frailty models are likelihood based: EM-algorithm (Klein, 1992), penalized partial likelihood (Therneau and Grambsch, 2000; McGilchrist, 1993), Bayesian analysis (Ducrocq and Casella, 1996). In recent papers more complex frailty models have been studied. Within the clinical trials context typical examples are frailty models with a random center effect and a random treatment by center interaction. To fit such frailty models, the likelihood based methods mentioned above have been adapted to cover this extra complexity in the data: EM algorithm (Vaida and Xu, 2000; Cortinas and Burzykowski, 2005), penalized partial likelihood (Ripatti and Palmgren, 2000),

Bayesian approach (Legrand *et al.*, 2005). We propose an alternative way to fit frailty models. We start from the following observation: the integral of the weighted (over time) conditional cumulative loghazard depends in a linear way on the random effects describing the cluster and/or the interaction heterogeneity and on the factor levels and/or covariates. Using the data within a cluster we can estimate the integral using nonparametric estimation techniques. Considering the estimated integral as a response we can reformulate the original problem of 'fitting a frailty model' into the problem of 'fitting a mixed model'.

67.2 From Frailty Model to Mixed Model

67.2.1 Model formulation

We consider clustered survival data with K different centers, center i having n_i patients. For each patient we observe the minimum of a failure time T_{ij}^0 and a right censoring time C_{ij} independent of T_{ij}^0; as notation we use $T_{ij} = \min(T_{ij}^0, C_{ij})$ for the observed time and δ_{ij} for the censoring indicator which is equal to 1 if $T_{ij} = T_{ij}^0$ and 0 otherwise. For each patient, we also have the binary variable x_{ij} representing the treatment to which the patient has been randomized with $x_{ij} = 0$ if the patient is in the standard arm and $x_{ij} = 1$ if the patient is in the experimental arm.
The following mixed-effects proportional hazards model is considered:

$$\lambda_{ij}(t) = \lambda_0(t) \exp\left(b_{0i} + (\beta + b_{1i})x_{ij}\right), \tag{67.2.1}$$

where $\lambda_0(t)$ represents the unspecified baseline hazard at time t, β is the fixed overall treatment effect, b_{0i} is the random center effect and b_{1i} is the random interaction effect providing information on how the treatment effect within center i deviates from the overall treatment effect captured by the regression coefficient β. The random effects b_{0i} and b_{1i} are assumed to follow zero-mean normal distributions. The variance-covariance matrix of the vector of random effects $\mathbf{b}^T = (b_{01}, b_{11}, \ldots, b_{0i}, b_{1i}, \ldots, b_{0K}, b_{1K})$ takes the form

$$\mathbf{G} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \otimes \mathbf{I}_K. \tag{67.2.2}$$

In absence of a random treatment by center interaction and in absence of covariates, model (67.2.1) reduces to the shared frailty model

$$\lambda_i(t) = \lambda_0(t) \exp(b_{0i}). \tag{67.2.3}$$

In (67.2.3) b_{0i}, $i = 1, \ldots, K$, is a sample from a zero-mean normal density with variance σ_0^2, describing the heterogeneity between centers.

67.2.2 The transformation

With $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(s)ds$ the cumulative hazard for the jth patient in center i, $j = 1, \ldots, n_i$ and $i = 1, \ldots, K$, and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, we easily obtain from (67.2.1) that

$$\ln \Lambda_{ij}(t) = \ln \Lambda_0(t) + b_{0i} + (\beta + b_{1i})x_{ij}. \tag{67.2.4}$$

Let $w(.)$ be a weight function $\left(W(t) = \int_0^t w(s)ds\right)$ satisfying $w(s) \geq 0, s \in [0, \infty)$ and $\int_0^\infty w(s)ds = 1$. Integrating both sides in (67.2.4) with respect to the weight function we obtain

$$\Omega_{ij} = \int_0^\infty \ln \Lambda_{ij}(t)dW(t) \quad = \quad \alpha + b_{0i} + (\beta + b_{1i})x_{ij},$$

with $\alpha = \int_0^\infty \ln \Lambda_0(t)dW(t)$. Since the patients in center i are divided, by the binary covariate x_{ij}, in a control and a treatment group we have that $\Omega_{i0} = \alpha + b_{0i}$ (control) and $\Omega_{i1} = \alpha + b_{0i} + (\beta + b_{1i})$ (treated). We also have that, for $k = 0, 1$,

$$\Omega_{ik} = \int_0^\infty \ln \Lambda_{ik}(t)dW(t)$$

with $\Lambda_{i0}(.)$, respectively $\Lambda_{i1}(.)$, the cumulative hazard function shared by all control, resp. treated, patients in group i. Following ideas in Grigoletto and Akritas (1999) pseudo observations for the Ω_{ik}'s can be obtained as

$$\hat{\Omega}_{ik} = \int_0^\infty \ln \hat{\Lambda}_{ik}(t)dW(t)$$

where $\hat{\Lambda}_{ik}(.)$ is the estimated cumulative hazard based on the observations (T_{ij}, δ_{ij}) for all patients in center i with, for $k = 0$, $x_{ij} = 0$ and, for $k = 1$, $x_{ij} = 1$. As concrete estimator we use $\hat{\Lambda}_{ik}(t) = -\ln \hat{S}_{ik}(t)$ with $\hat{S}_{ik}(t)$ the Kaplan-Meier estimator.
In terms of the pseudo observations we now can propose the model

$$\hat{\Omega}_{ik} = \alpha + b_{0i} + (\beta + b_{1i})x_{ik} + (\hat{\Omega}_{ik} - \Omega_{ik}) = \alpha + b_{0i} + (\beta + b_{1i})x_{ik} + e_{ik} \tag{67.2.5}$$

with $x_{i0} = 0$ and $x_{i1} = 1$. Since $e_{ik} = \hat{\Omega}_{ik} - \Omega_{ik}$ it is clear that the random error terms do not satisfy the heterogeneity assumption (since different subclusters have different sample sizes). Based on a stochastic approximation we can obtain (asymptotic) i.i.d. representations for the error terms and, hence, we obtain an explicit expression for the variance of e_{ik}. By using the estimated variances of the error terms, we account for the heterogeneity of the error terms when mixed model software is used to fit the model.
For the special case (67.2.3) we obtain the following model after transformation:

$$\hat{\Omega}_i = \alpha + b_{0i} + \left(\hat{\Omega}_i - \Omega_i\right) = \alpha + b_{0i} + e_i. \tag{67.2.6}$$

For this one-way random effects model we only have one observation per center. At first glance this leads to identifiability problems. We, however, do have estimators of the variances of the error terms so that estimation of the variance components associated with the random center effect is possible.

67.3 Simulations

We investigate the performance of the proposed method based on a simulation study. As a simulation model we consider the setting of a multicenter clinical trial with treatment as a covariate. First, we consider the special case in model (67.2.3) where there is only a random center effect. We compare the results obtained by the proposed method with those obtained by the penalized partial likelihood approach. We discuss the precision of the parameter estimates for the varying number of clusters and the number of observations per cluster, the percentage of censored observations, the size of σ_0^2 and the value of the baseline event rate λ_0 (which we assume constant in time for simplicity). The results indicate that σ_0^2 is estimated well by the proposed method if the cluster size is large enough. Both for the penalized partial likelihood approach and the proposed method, the absolute relative bias decreases with increasing cluster size but is not substantially influenced by the number of clusters. In general, the absolute relative bias increases if the amount of censoring increases. Further, the results illustrate that the point estimates of σ_0^2 are biased if the frailty distribution is misspecified. This is a problem for both methods (see also Massonnet *et al.*, 2006a). Next, the general model (67.2.1) is considered. For this model, we allow for correlation between b_{0i} and b_{1i}. The effect of the size of σ_0^2 and σ_1^2 on the precision of the parameter estimates is discussed. The simulations show a good performance of the proposed method.

67.4 Conclusions

We propose an alternative approach to fit frailty models. Based on the original data we obtain pseudo-data (the estimated integrals) on which we can apply mixed model theory. The simulation study illustrates that the proposed method provides a good and simple alternative for fitting frailty models for data sets with a sufficiently large number of clusters and moderate to large sample sizes within clusters.

Acknowledgment

The authors gratefully acknowledge the financial support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

References

1. Cortinas Abrahantes, J., Burzykowski, T. (2005). A version of the EM algorithm for proportional hazards model with random effects, *Biometrical Journal*, **47**, 847–862.

2. Ducrocq, V., Casella, G. (1996). A Bayesian analysis of mixed survival models, *Genetics Selection Evolution*, **28**, 505–529.

3. Grigoletto, M., Akritas, M.G. (1999). Analysis of covariance with incomplete data via semiparametric model transformations, *Biometrics*, **55**, 1177–1187.

4. Legrand, C., Ducrocq, V., Janssen, P., Sylvester, R., Duchateau, L. (2005). A Bayesian approach to jointly estimate center and treatment by center heterogeneity in a proportional hazards model, *Statistics in Medicine*, **24**, 3789–3804.

5. Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm, *Biometrics*, **48**, 795–806.

6. Massonnet, G., Burzykowski, T., Janssen, P. (2006a). Resampling plans for frailty models, *Communications in Statistics - Simulation and Computation*, **35**, To appear in May 2006.

7. Massonnet, G., Janssen, P., Burzykowski, T. (2006b). Fitting frailty models via linear mixed models using model transformation, Technical Report.

8. McGilchrist, C. (1993). REML estimation for survival models with frailty, *Biometrics*, **49**, 221–225.

9. Ripatti, S., Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics*, **56**, 1016–1022.

10. Therneau, T.M., Grambsch, P.M. (2000). *Modeling Survival Data, Extending the Cox model*, Springer, New York.

11. Vaida, F., Xu, R. (2000). Proportional hazards model with random effects, *Statistics in Medicine*, **19**, 3309–3324.

68

On Measures of Divergence and the Divergence Information Selection Criterion

Kyriacos Mattheou and Alex Karagrigoriou

Department of Mathematics and Statistics, University of Cyprus, Cyprus

Abstract: The aim of this work is to develop a new model selection criterion using a general discrepancy based technique, by constructing an asymptotically unbiased estimator of the overall average discrepancy between the true and the fitted models. Furthermore, the lower bound for the mean squared error of prediction is established.

Keywords and phrases: DIC, power divergence, MSE of prediction

68.1 Introduction

The divergence measures are used as indices of similarity or dissimilarity between populations. They are also used either to measure mutual information concerning two variables or to construct model selection criteria. A model selection criterion can be constructed as an approximately unbiased estimator of an expected "overall discrepancy" (or divergence), a nonnegative quantity which measures the "distance" between the true model and a fitted approximating model. A well known discrepancy is Kullback-Leibler discrepancy that was used by Akaike (1973) to develop Akaike Information Criterion (AIC).

Measures of discrepancy or divergence between two probability distributions have a long history. A unified analysis was recently provided by Cressie and Read (1984) who introduced for both the continuous and the discrete case the so called power divergence family of statistics that depends on a parameter λ and is used for multinomial goodness-of-fit tests. The additive and non-additive directed divergences of order α were introduced in the 60's and the 70's (Renyi, 1961 and Rathie and Kannappan, 1972). It should be noted that for λ tending to 0 and for α tending to 1 the above measures become the Kullback-Leibler measure. Another family of measures is the Φ-divergence known also as

Csiszar's measure of information (Csiszar, 1963) the discrete form of which is given by $I^c(P;Q) = \sum_{i=1}^{k} q_i \Phi(p_i/q_i)$, where Φ is a real valued convex function on $[0,\infty]$ and $P = (p_1, p_2, \ldots, p_k)$ and $Q = (q_1, q_2, \ldots, q_k)$ are two discrete finite probability distributions. For various functions for Φ the measure takes different forms. The Kullback-Leibler measure is obtained for $\Phi(u) = u\log(u)$ while the additive directed divergence is obtained for $\Phi(u) = sgn(\alpha - 1)u^{\alpha}$ and for the transformation $(\alpha - 1)^{-1}\log I^c$. For a comprehensive discussion on measures of divergence the reader is referred to Pardo (2006).

A new discrepancy measure was recently introduced by Basu et. al (1998). In this paper, we develop a new model selection criterion which is an approximately unbiased estimator of the expected overall power divergence that corresponds to Basu's power divergence measure. Furthermore, we obtain a lower bound for the mean squared error (MSE) of prediction.

68.2 Basu's Power Divergence Measure

One of the most recently proposed discrepancies is Basu's Power Divergence [Basu et. al (1998)] which is defined as:

$$d_a(g,f) = \int \left\{ f^{1+a}(z) - \left(1 + \frac{1}{a}\right) g\,(z)\, f^{a}\,(z) + \frac{1}{a} g^{1+a}\,(z) \right\} dz,\ a > 0 \quad (68.2.1)$$

where g is the true model, f the fitted approximating model, and a a positive number. The discrete form of the measure is given by

$$\sum_{i=1}^{k} \left\{ p_i^{1+a} - (1 + \frac{1}{a}) p_i^a q_i + \frac{1}{a} q_i^{1+a} \right\},$$

where p_i and q_i, $i = 1, 2, \ldots, k$ are as in Section 1.1. Observe that the above measure takes the form $\sum_{i=1}^{k} q_i^{1+a} \Phi(p_i/q_i)$ where $\Phi(u) = u^{1+a} - (1 + a^{-1})u^a + a^{-1}$.

Lemma 68.2.1 *The limit of (68.2.1) when $a \to 0$ is the Kullback-Leibler divergence. Furthermore, the discrete form of the measure tends to the Kullback-Leibler measure for $a \to 0$ and for $\Phi(u) = u\log u$.*

It is easy to see that Basu's measure satisfies the basic properties of measures, namely the properties of nonnegativity and the continuity. In particular, the value of measure is nonnegative while small changes in the distributions result in small changes in the measure. Finally, the value of the discrete measure is not affected by the simultaneous and equivalent reordering of the discrete masses which confirms the symmetry property of the Basu's measure.

Consider a random sample $X_1, \ldots, X_n$ from the distribution g and a candidate model f_t from a parametric family of models $\{f_t\}$, indexed by an unknown

parameter $t \in \Theta$. To construct the new criterion for goodness of fit we shall consider the quantity:

$$W_t = \int \left\{ f_t^{1+a}(z) - \left(1 + \frac{1}{a}\right) g(z) f_t^a(z) \right\} dz, \ a > 0. \tag{68.2.2}$$

which is the same as (68.2.1) without the last term that remains constant irrespectively of the model f_t used. Observe that (68.2.2) can also be written as:

$$W_t = \int f_t^{1+a}(z)\, dz - \left(1 + \frac{1}{a}\right) E_g\left(f_t^a(z)\right), \ a > 0. \tag{68.2.3}$$

Our target theoretical quantity that would be estimated by the new criterion is

$$E\left(W_t \left| t = \hat{\theta}\right.\right) \tag{68.2.4}$$

which can be viewed as the average distance between g and f_t up to a constant and is known as the expected overall discrepancy between g and f_t. In (68.2.4), $\hat{\theta}$ is the estimator of t that minimizes an estimate of $d_a(g, f_t)$ with respect to t. Note that the estimator of t is obtained by minimizing (68.2.3) when the expectation is replaced by its sample analogue, namely $n^{-1}\sum_{i=1}^n f_t^a(X_i)$. In the theorem below, Basu et. al. (1998) provide the asymptotic properties of $\hat{\theta}$.

Theorem 68.2.1 (Basu et. al (1998)) *Under regularity conditions, there exists estimator $\hat{\theta}$ which is consistent and asymptotically normal with mean zero and variance $J(\theta)^{-2}K(\theta)$, where under the assumption that the true distribution g belongs to the parametric family $\{f_t\}$, θ being the true value of the parameter and $\xi = \int u_\theta(z) f_\theta^{1+a}(z)\, dz$,*

$$J(\theta) = \int [u_\theta(z)]^2 f_\theta^{1+a}(z)\, dz \quad \text{and} \quad K(\theta) = \int [u_\theta(z)]^2 f_\theta^{1+2a}(z)\, dz - \xi^2. \tag{68.2.5}$$

The Lemma below provides the derivatives of (68.2.3) in the case where g belongs to the family $\{f_t\}$ (see Mattheou and Karagrigoriou (2006a)).

Lemma 68.2.2 *For $a > 0$ and if the true distribution g belongs to the parametric family $\{f_t\}$, the derivatives of (68.2.3) are:*

$$(a)\ \frac{\partial W_t}{\partial t} = (a+1)\left[\int u_t(z) f_t^{1+a}(z)\, dz - E_g\left(u_t(z) f_t^a(z)\right)\right] = 0,$$

$$(b)\ \frac{\partial^2 W_t}{\partial t^2} = (a+1)\left\{(a+1)\int [u_t(z)]^2 f_t^{1+a}(z)\, dz - \int i_t f_t^{1+a} dz \right.$$
$$\left. + E_g\left(i_t(z) f_t^a(z)\right) - E_g\left(a[u_t(z)]^2 f_t^a(z)\right)\right\} = (a+1)J$$

where $u_t = \frac{\partial}{\partial t}(\log(f_t))$, $i_t = -\frac{\partial^2}{\partial t^2}(\log(f_t))$ and $J = \int [u_t(z)]^2 f_t^{1+a}(z)\, dz$.

Theorem 68.2.2 *Under the assumptions of Lemma (68.2.2) and for a p-dimensional parameter t, the expected overall discrepancy at* $t = \hat{\theta}$ *is given by*

$$E\left(W_t|\, t=\hat{\theta}\right) = W_\theta + \frac{(a+1)}{2} E\left[\left(\hat{\theta}-\theta\right) J \left(\hat{\theta}-\theta\right)'\right]. \tag{68.2.6}$$

68.3 The divergence information criterion

In this section we introduce the new criterion and prove that it is an approximately unbiased estimator of (68.2.4). Due to the unknown true distribution g, we estimate (68.2.3) by the empirical distribution function given by:

$$Q_t = \int f_t^{1+a}(z)\, dz - \left(1+\frac{1}{a}\right)\frac{1}{n}\sum_{i=1}^{n} f_t^a(X_i). \tag{68.3.7}$$

Lemma 68.3.1 *The derivatives of (68.3.7) are:*

$$(a)\ \frac{\partial Q_t}{\partial t} = (a+1)\left[\int u_t(z) f_t^{1+a}(z)\, dz - \frac{1}{n}\sum_{i=1}^{n} u_t(X_i) f_t^a(X_i)\right],\ a>0,$$

$$(b)\ \frac{\partial^2 Q_t}{\partial t^2} = (a+1)\left\{(a+1)\int [u_t(z)]^2 f_t^{1+a}(z)\, dz - \int i_t f_t^{1+a}(z)\, dz \right.$$
$$\left. + \frac{1}{n}\sum_{i=1}^{n} i_t(z) f_t^a(z) - \frac{1}{n}\sum_{i=1}^{n} a\,[u_t(z)]^2 f_t^a(z)\right\}$$

where $u_t = \frac{\partial}{\partial t}(\log(f_t))$ *and* $i_t = -(u_t)' = -\frac{\partial^2}{\partial t^2}(\log(f_t))$.

It is easy to see that by the weak law of large numbers, as $n \to \infty$, we have:

$$\left[\frac{\partial Q_t}{\partial t}\right]_\theta \to_P \left[\frac{\partial W_t}{\partial t}\right]_\theta \quad \text{and} \quad \left[\frac{\partial^2 Q_t}{\partial t^2}\right]_\theta \to_P \left[\frac{\partial^2 W_t}{\partial t^2}\right]_\theta. \tag{68.3.8}$$

The consistency of $\hat{\theta}$ and (68.3.8) can be used to evaluate the expectation of the empirical estimator evaluated at the true point θ.

Theorem 68.3.1 *Under the assumptions of Lemma (68.2.2), the expectation of* Q_t *evaluated at* θ *is given by*

$$EQ_\theta \equiv E(Q_t|t=\theta) = EQ_{\hat{\theta}} + \frac{a+1}{2} E\left[\left(\theta-\hat{\theta}\right) J \left(\theta-\hat{\theta}\right)'\right].$$

The asymptotically unbiased estimator of $E\left(W_t \left| t=\hat{\theta}\right.\right)$ is provided in the theorem below (see Mattheou and Karagrigoriou (2006b)).

Theorem 68.3.2 *An asymptotically unbiased estimator of the expected overall discrepancy evaluated at $\hat{\theta}$ is given by*

$$DIC = Q_{\hat{\theta}} + (a+1)(2\pi)^{-\frac{a}{2}} \left(\frac{1+a}{1+2a}\right)^{1+\frac{p}{2}} p. \tag{68.3.9}$$

68.4 Lower bound of the MSE of prediction

Let Xj be the design matrix of the model $Y = X_j\beta + \varepsilon$ where $\beta = (\beta_0, \beta_1, \beta_2, \ldots)'$, $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$ and I is the infinite dimensional identity matrix.

Let $V(j) = \{\beta(j),\ \text{s.t.}\ \ \beta(j) = \left(\beta_0, 0, ..., \beta_{j_1}, 0, ..., \beta_{j_{k_j}}, 0, ...\right)\}$ be the subspace that contains only the $k_j + 1$ parameters β_{j_i} involved in the model and let $\beta^{(n)}$ to be the projection of β on $V(j)$.

The prediction $\hat{y}$ is given by $\hat{y} = X_j\hat{\beta}$, where the estimator of $\beta^{(n)}$ obtained through a set of observations $(X_{ij_1}, \ldots, X_{ij_{k_j}}, y_i)$, $i = 1, 2, \ldots, n$ is denoted by $\hat{\beta} = \left(\hat{\beta}_0, 0, ..., \hat{\beta}_{j_1}, 0, ..., \hat{\beta}_{j_2}, 0, ..., \hat{\beta}_{j_{k_j}}, 0, ...\right)'$.

The mean squared error (MSE) of prediction and the average MSE of prediction are defined respectively by

$$Q_n(j) = \mathrm{E}\left[(\hat{y}_{n+1} - y_{n+1})^2 | X\right] - n\sigma^2 \ \text{ and } \ L_n(j) \equiv \mathrm{E}\left(Q_n(j)\right).$$

Lemma 68.4.1 *Under the notation and conditions of this section we have that*

$$Q_n(j) = \left\|\hat{\beta} - \beta\right\|^2_{M_n(j)} \quad \textit{and} \quad L_n(j) = \mathrm{E}\left\|\hat{\beta} - \beta\right\|^2_{M_n(j)},$$

where $M_n(j) = X_j'X_j$ and $\|A\|^2_R = A'RA$.

The Lemma below provides a lower bound for the MSE of prediction. In particular, we show that $Q_n(j)$ is asymptotically never below the quantity $L_n(j^*) = \min_j L_n(j)$.

Lemma 68.4.2 *Let $L_n(j^*) = \min_j L_n(j)$. Under certain regularity conditions, we have that for every $\delta > 0$*

$$\lim_{n\to\infty} P\left[\frac{Q_n(j)}{L_n(j^*)} > 1 - \delta\right] = 1.$$

68.5 Discussion

Note that the family of candidate models is indexed by a single parameter a. The value of a dictates to what extent the estimating methods become more robust than the maximum likelihood methods. One should be aware of the fact

that the larger the value of a the bigger the efficiency loss. As a result one should be interested in small values of $a \geq 0$, say between zero and one.

The proposed DIC criterion could be used in applications where outliers or contaminated observations are involved. The prior knowledge of contamination may be useful in identifying an appropriate value of α. Preliminary simulations with a 10% contamination proportion show that DIC has a tendency of underestimation in contrast with AIC which overestimates the true model.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proc. of the* 2^{nd} *Intern. Symposium on Information Theory*, (Petrov B. N. and Csaki F., eds.), Akademiai Kaido, Budapest.

2. Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence, *Biometrika*, **85**, 549-559.

3. Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. R. Statist. Soc*, **5**, 440–454.

4. Csiszar, I. (1963). Eine informationstheoretische ungleichung and ihre anwendung auf den beweis der ergodizitat von markoffischen ketten. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, **8**, 85-108.

5. Mattheou, K. and Karagrigoriou, A. (2006a). A discrepancy based model selection criterion, *Proc. 18th Conf. Greek Statist. Society* (to appear).

6. Mattheou, K. and Karagrigoriou, A. (2006b). On asymptotic properties of DIC, *TR 09/2006*, Dept. of Math. and Stat., Univ. of Cyprus.

7. Pardo, L. (2006). *Statistical inference based on divergence measures*, Chapman and Hall/CRC.

8. Rathie P.N. and Kannappan, P. (1972). A directed-divergence function of type β. *Information and Control*, **20**, 38-45.

9. Renyi, A. (1961). On measures of entropy and information. *Proc.* 4^{th} *Berkeley Symp. on Math. Statist. Prob.*, **1**, 547-561, Univ. of California Press.

69

Application of Inverse Problems in Epidemiology and Demography

A. Michalski

Institute of control sciences, Moscow

Abstract: Different problems in epidemiology and demography can be considered as solutions of inverse problems, when using observed data one estimates the process caused the data. Examples are estimation of infection rate on dynamics of the disease, estimation of mortality rate on sample of survival times, estimation of survival in wild on survival in laboratory. A specific property of inverse problem - instability of solution is discussed, the procedure for stabilization is presented. Examples of morbidity estimation on incomplete data, HIV infection rate estimation on dynamics of AIDS cases and estimation of survival function in wild population on survival of captured animals are presented.

Keywords and phrases: Inverse problem, Epidemiology, Demography, Incomplete follow-up, HIV infection rate, AIDS cases dynamics, Survival in wild

69.1 Introduction

Interpretation of observations in different disciplines of life sciences can be considered as a solution of mathematical inverse problem. Examples are epidemiology, demography and biodemography. The important in epidemiology indicators such as prevalence of a disease and incidence of it are related by cause-effect relationship. This means that the process of a disease occurrence in formal way causes process of accumulation of the disease cases in population. The other example is relationship between rate of infection and the number of corresponding diagnosed cases. In demography cause-effect relationship exists between mortality and survival processes. Mortality process 'make influence' on survival in population.

In all these examples the value of the 'effect' can be estimated on population observations while the direct estimation of the 'cause' is impossible or

needs great funds. On the other hand information about the 'cause' often is important for better understanding of the phenomenon investigated and mathematical methods for estimation of 'cause' on 'effect' data are needed. The report describes mathematical formulations of the 'cause-effect' problem, difficulties of estimation of the 'cause process' on population data and a procedure elaborated to overwhelm them. Three examples with results of calculations are presented: estimation of morbidity on the results of incomplete follow-up, estimation of HIV infection rate on the dynamics of AIDS cases, estimation of survival in wild population on survival of captured animals in laboratory.

69.2 Mathematical formulation

Many problems from epidemiology and demography can be written as a relationship between an unobserved process $\Psi(x)$ and an observed process $U(\mathrm{x})$ in form

$$U(x) = A\Psi = \int_a^b K(x,t)\Psi(t)dt, \tag{69.2.1}$$

where A is the integral operator given by a kernel function $K(x,t)$, which is defined by the nature of the problem investigated. More detailed consideration for this function is given bellow. The main property of the equation (69.2.1) with continues kernel is that exact solution is unstable in respect to small variations in the observed function $U(\mathrm{x})$. In mathematical terms this means that there exists a sequence of functions δU such that the sequence of corresponding solutions of equation (69.2.1) $\delta\Psi = A^{-1}\delta U$ do not tend to zero while the sequence δU tends to zero. Such problems are called ill-posed problems by Tikhonov and Arsenin (1977). In practical applications this means that a small disturbance in observations U can lead to big disturbance in the exact solution $A^{-1}U$. Such property is well known in numerical solution of large sets of linear equations. In this case A is a matrix such that matrix A^TA has small eigenvalue, which means that the inverse matrix $\left(A^TA\right)^{-1}$ has large eigenvalue and disturbance in the solution $\Psi = \left(A^TA\right)^{-1}A^TU$ is high. Often the sensitivity of the system is so high that even machine arithmetic errors are enough to change the solution Ψ dramatically.

To obtain the stable solution for the equation (69.2.1) one is to put additional restrictions to the solution. Tikhonov and Arsenin (1977) proposed to put such restrictions by minimizing on Ψ a functional

$$\|U - A\Psi\|^2 + \alpha\Omega\left(\Psi\right), \tag{69.2.2}$$

where $\Omega\left(.\right) > 0$ is a stabilization functional, defined such that for any constant C the set $\{\Psi : \Omega(\Psi) \leq C\}$ is a compact set and α is a positive stabilization

parameter. The optimal value for α depends on the level of disturbance δ in observed data U. It is proved that if $\delta^2/\alpha \to 0$ while $\delta \to 0$ and $\alpha \to 0$, then the minimizer of (69.2.2) Ψ_α tends to the exact solution of equation (69.2.1). The problem of proper selection value for the stabilization parameter α if the level of disturbance δ does not tend to zero is still a challenging task. Different approaches and methods are described in Evans and Stark (2002) including cross-validation and Bayesian approaches.

The different approach to stabilization parameter selection is based on estimate for mathematical expectation for quadratic functional value minimized on finite sample, which is described in Michalski (1987). For the solution Ψ_α, which minimizes the functional $\|U - A\Psi\|^2 + \alpha \|B\Psi\|^2$ for α such that $\mathrm{m} > 2TrA_\alpha$, with probability no less than 1-η the inequality is valid

$$E_{Y,U} \|Y - A\Psi_\alpha\|^2 < \frac{\|U - A\Psi_\alpha\|^2}{1 - 2TrA_\alpha/\mathrm{m}} + const + \sqrt{\frac{const}{\eta}}. \tag{69.2.3}$$

Here U, Y are independent realizations of size m, generated from the same distribution, A, B are matrices and $A_\alpha = A\left(A^TA + \alpha B^TB\right)^{-1}A^T$. The left hand side of (69.2.3) is the mean value of disagreement between possible vectors of experimental and predicted data. To get it small one can use for stabilization parameter α value, which minimizes $I_\alpha = \|U - A\Psi_\alpha\|^2 / (1 - 2TrA_\alpha/\mathrm{m})$. The quantity $\|U - A\Psi_\alpha\|^2$ is a square residual for empirical data.

It is interesting to note that the cross-validation criterion takes the form $I_\alpha^{cv} = \|U - A\Psi_\alpha\|^2 / (1 - TrA_\alpha/\mathrm{m})^2$. For a small amount of data it is demonstrated in Michalski (1987), that the criterion I_α produces better results than the cross-validation criterion I_α^{cv}.

69.3 Estimation of morbidity on the results of incomplete follow-up

Michalski *et al.* (1996) considered a problem of estimation morbidity on data of irregular health examinations. This problem leads to solution of a matrix equation

$$A\Psi = U$$

with U the proportion of diagnosed cases among observed people by years of investigation and Ψ the probability for healthy person to become sick by years. A is a triangular matrix with 1 at the main diagonal and elements a_{ij} equal to the proportion of people, examined in the year i and been healthy before, among those, who skipped the examination in the year j after the last examination. In the case of a complete follow-up study the matrix A is the identity matrix and

the morbidity estimate for different years is just the ratio between the number of cases and the number of people, examined in the same year.

Stabilization of the matrix equation was made by minimization (69.2.2) with the stabilization functional $\Omega(p) = \|B\Psi\|^2 = \Psi^T B^T B\Psi$ and B the matrix with two non zero diagonals. It holds -1 at the main diagonal and 1 at the second. This structure of stabilization functional reflects the hypothesis, that the morbidity will not change significantly in consequent years.

The described approach was applied in Michalski *et al.* (1996) for the estimation of malignant neoplasm (ICD9 140-208) morbidity among participants in the clean-up operations after the accident on the Chernobyl Nuclear Power Station in 1986. The value for stabilization parameter α was selected using the above described criterion I_α. Estimates show that observed morbidity increases in time with higher rate than the real, unobserved one, because of 'morbidity accumulation' effect among people skipping regular examinations. The described approach adjusts estimates for this effect.

69.4 Estimation of HIV infection rate on the dynamics of AIDS cases

Large latent period between HIV infection an AIDS manifestation makes it difficult to judge about the amount of HIV infected people in population. Specific expensive surveys of risk groups are needed to get reliable information about HIV prevalence. Implementation of inverse problems approach can help to estimate the number of HIV infected people from dynamics of AIDS cases, which is reported for the health care system needs. The number of people infected by HIV in year t at age x $\Psi(t,x)$ is related with the number of AIDS diagnoses in year t at age x $U(t,x)$ by the integral equation

$$U(t,x) = \int_0^x L(x,s) \exp\left(- \int_s^x \mu_c(t-x+\tau,\tau)\, d\tau \right) \Psi(t-x+s,s)\, ds \tag{69.4.4}$$

where $\mu_c(t,x)$ is the mortality in year t at age x, $L(x,s)$ is the probability density function for the distribution of AIDS diagnoses age x if at age s a person was infected with HIV. Age specific mortality supposed to be known from national data and the function $L(x,s)$ can be estimated from the clinical data and data about AIDS cases among patients which were infected with HIV during blood transfusion. The most common models for $L(x,s)$ are exponential, Weibull and Markov chain model. Write equation (69.4.4) in matrix form

$$U = A\Psi,$$

where U and Ψ are vectors composed by values of functions $U(.)$ and $\Psi(.)$ for the corresponding birth cohorts and A is the block-diagonal matrix composed by triangular matrices with elements for k-th cohort

$$a_{ij}^k = \begin{cases} 0 & s_j > x_i^k \\ \beta(t_i^k, x_i^k) L(x_i^k, s_j) \exp\left(- \int\limits_{s_j}^{x_i^k} \mu_c \left(d_k + \tau, \tau\right) d\tau \right) & s_j \leq x_i^k \end{cases} .$$

To stabilize the solution of (69.4.4) the stabilization functional was used in the form

$$\|Y - A\Psi\|^2 + \alpha\Omega\left(\Psi\right)$$

with $\Omega\left(\Psi\right) = \sum\limits_k \frac{1}{m_k} \sum\limits_{j=2}^{m_k} \left(\Psi_j^k - \Psi_{j-1}^k\right)^2$. The stabilized solution takes the form

$$\Psi_\alpha = \left(A^T A + \alpha D\right)^{-1} A^T Y,$$

where the matrix D is a block-diagonal matrix composed by three diagonal matrices. For the *k-th* cohort the matrix holds $2/m_k$ at the main diagonal, $-1/m_k$ at the other two diagonals and $1/m_k$ as the first and the last elements of the matrix.

Results of HIV infection rate from AIDS diagnoses dynamics estimation on simulated data are presented. The stabilization parameter value was selected using the described criterion I_α.

69.5 Estimation of survival in wild population on survival of captured animals in laboratory

A specific problem arises in connection with investigation of life span in wild populations of different species. The problem of how to estimate the survival curve in wild population of flies is considered in Muller *et al.* (2004). A portion of wild flies were cached and kept in laboratory in conditions similar to conditions in wild nature. The survival curve was calculated for cached cohort and some mathematical technique is to be applied to produce the survival curve for wild population. This is a typical inverse problem. If laboratory conditions do not change, survival of fly then survival in laboratory $S_{lab}(.)$ is related with survival in wild stable population $S_{wild}(.)$ by the integral equation

$$S_{lab}(x) = \frac{1}{e_0} \int\limits_x^\omega S_{wild}(y) dy \tag{69.5.5}$$

where ω is the maximum life span and e_0 is the life expectancy in wild population. By differentiating the last equation on x one obtains the equation

$$S_{wild}(x) = \frac{\frac{d}{dx}S_{lab}(x)}{\frac{d}{dx}S_{lab}(0)}.$$

One can estimate numerically the derivative from the survival function in laboratory and calculate from it $S_{wild}(x)$. This is done in Muller *et al.* (2004).

The other possibility is to solve numerically equation (69.5.5) itself. The corresponding matrix equation is

$$AX = S_l \tag{69.5.6}$$

where $X = \frac{1}{e_0}S_w$, S_w and S_l are vectors of values of survival functions observed daily in wild and in laboratory populations respectively and A is a triangular matrix with 0 bellow the main diagonal and 1 at the other places. Solution of the system (69.5.6) was stabilized as described above. Results of estimation with simulated and real data are presented. The stabilization parameter value was selected using the described criterion I_α.

References

1. Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*, Wiley, New York.

2. Evans, S. N. and Stark, P. N. (2002). Inverse problems as statistics, *Inverse Problems*, **18**, R55-R97.

3. Michalski A. I. (1987). Choosing an algorithm of estimation based on samples of limited size, *Automatization and Remote Control*, **48**, 909-918.

4. Michalski, A.I., Morgenstern, W., Ivanov, V.K. and Maksyitov, M.A. (1996). Estimation of morbidity dynamics from incomplete follow-up studies, *Journal Epidemiology and Biostatistics*, **1**, 151-157.

5. Muller, H.-G., Wang, J.-L., Carey, J. R., Caswell-Chen, E. P., Chen, C., Papadopoulos, N. and Yao, F. (2004). Demographic window to aging in the wild: constructing life tables and estimating survival functions from marked individuals of unknown age, *Aging Cell*, **3**, 125-131.

70

Neglected Issues In The Application Of Statistics To Epidemiology And Medicine

Minder, Ch. E

Dept of Social and Preventive Medicine, University of Berne, Switzerland

Abstract: In this paper, the view is presented that statistics has evolved to such an extent in the last decades that there are new areas in need of attention, while traditionally central issues may have lost some of their former urgency. For decades, there was an overwhelming need to develop ever more general, refined and complex models. Considering the application of statistics, many problems seem to result from an inadequate understanding and implementation of statistical models by their (non-specialist) users. Unreflected, schematic, and hence inadequate application of well developed and statistically well investigated models appear to be frequent. Moreover, little reference is made to checking basic model assumptions. Indeed, it is often unclear whether sufficient checks were performed, if any at all. In applications of statistics in epidemiology and medicine, comparison with earlier results is mostly limited to comparing the magnitude of certain isolated coefficients, while the full reporting of the models applied is rare.

This paper raises these issues with the aim of initiating a discussion which may hopefully lead in time to improvement of the situation.

Keywords and phrases: Applied, assumption, checking, full model, modelling, quality, reporting

70.1 Introduction

This paper is not about a new technique or a new application of a known technique, nor will it present new theorems or proofs. Its aim is rather to bring to attention some often forgotten points in the application of statistics in science. This is done at the hand of epidemiology and occasionally medicine. The points raised are selected based on the experiences of the author in these

fields of application. However, they likely apply more or less also to other fields of application of statistics.

The following deliberations apply only to statistical models used to make inferences about causes and effects. In statistical terms, they only concern models based on regression approaches. These comments do not apply to other procedures, such as multivariate correlational procedures treating all variables involved in a symmetric way.

70.2 Model building

One important difference of the application of statistical models in epidemiology as opposed to the exact sciences is the uncertainty about the appropriate model. While in science, the form of the statistical model to be used is frequently (not always, however!) given by theory, this is not so in fields like epidemiology and other biology-based and social sciences. Here, the model is generally much more ad hoc and, apart from offering the possibility of estimating possibly causal effects, one of it's main purposes is to provide a concise description of an otherwise confusing heap of data.

There is a tendency to apply refined statistical models to complex data sets in an uncritical manner. This trend is fuelled by the widespread availability of statistical software of stupendous power, flexibility and ease of use. The availability of such sophisticated software is, however, not balanced by sufficient efforts to educate the epidemiologist and medical users adequately in statistics. As a consequence, these users have only limited ability to appreciate the necessities and constraints of good modelling. Even simple facts, e.g. that the user needs to provide a well reflected predictor function are, in my experience, not well recognized. The situation is even worse with regard to knowledge about the assumptions underlying specific procedures. These appear to be known only on a most basic level. It is e.g mostly known that the scale of measurement of the dependent variable is relevant for the choice of a model. With regard to the limitations of any chosen model, knowledge appears to be very limited indeed.

The technology of "canned" statistical software is used by epidemiologists and medical researchers in much the same way we all use our cars, mobile phones and computers: without understanding the details of their functioning. This is risky behaviour in all fields. The risks with statistical techniques and software are special in that they do not only concern the users of statistical techniques, but also the well-being of patients, the reputation of epidemiology and medicine as sciences and of statistics as an academic discipline.

As a consequence of insufficient statistical training of epidemiologists, there appears to prevail a general lack of understanding on how to proceed in order to obtain a useful statistical model for a given epidemiological setting. Dummy variables seem to be handled mostly in a correct way. Despite extensive available

literature, most models with continuous covariates incorporate these either in a grouped form or linearly, whether appropriate or not. The consequence is a schematic, unimaginative and hence inappropriate, suboptimal use of statistical models.

70.3 Model checking

The foregoing section on model building illustrates the importance of model checking. If only little effort goes into selecting a suitable model, model checking becomes all the more important. Model checking is implemented in epidemiological research to the extent that it has become commonplace to run sensitivity analyses, using alternate models and/or excluding certain portions of the data. Also, there are often efforts to simplify the model, either by dropping insignificant variables, or by reporting them as such.

It is, in my experience, however, not commonplace to do extensive checks of the assumptions the model is based on (e.g. linearity, absence of collinearity, conformity of the residual variation with the form stipulated by the model, distributional assumptions).

With regard to examining linearity and taking into account collinearity, observational sciences such as epidemiology and econometrics are at a disadvantage, since intrinsic collinearities between important covariables pose an obstacle difficult to surmount. However, often neither the available data, nor the statistical models allow a realistic and checkable modelling of the underlying process.

In some instances, the absence of model checking is due to the rarity and lack of prominence of papers on appropriate methods. Thus, the annual plethora of new models is not matched by an equal abundance of model checking procedures.

The main obstacle to proper model checking is however, the lack of understanding of its importance on the part of researchers, authors, reviewers and editors of scientific journals.

70.4 Model reporting

Proper reporting of the model(s) used along with the data fitted is often lacking in epidemiology and medicine. An important reason in the past was lack of journal space, limiting authors to reporting only the very "sexiest" parts of their research. Here, the internet may (and already has) brought some relief.

The lack of information on the models used to reach conclusions in a publication is a significant handicap for combining the relevant information of several studies using the tools of meta-analysis.

The rise of meta-analysis has, however, had the beneficial effect of focussing the attention of clinical epidemiologists on the necessity of reporting standards. At the time these cover mainly procedural aspects of the studies to be re-analysed.

70.5 What can be done?

Probably the most effective way to promote better standards in model building, checking and reporting is by promoting the understanding and sensibility of the boards and editors of scientific journals to these issues.

This could be achieved by some reputable scientific body representing statistics setting up a "Committee on Good Statistical Practice". Duty of this Committee would be to devise a set of "state-of-the-art" guidelines on how to do model building, checking and reporting for scientific studies and in scientific reports.

To that end, some (statistical) research has to be undertaken in the corresponding journals documenting the statistical quality and shortcomings of these channels.

Another way of improving the statistical quality of epidemiological and medical publications is to encourage the employment of trained statisticians in research teams. Equally, involving trained statisticians (well versed in applying statistics to epidemiology or medicine) on a regular basis in the reviewing process would help too. A further-reaching proposal is to institute statistical peer review by requiring the presentation of data of any publication for re-analysis and by regular re-analysis. The implications of this last proposal must however be investigated.

70.6 Discussion

It is hoped that the paper illustrates that statisticians cannot possibly limit themselves to "do mathematics", develop ever new models and investigate the finer properties of well known procedures. Rather, a concerted effort has to be made at several fronts (a) to bring to proper use models and procedures developed by mathematical statisticians; (b) to intensify the research and publication efforts of mathematical statisticians on model checking procedures (it is this authors belief that model averaging will not be the solution); (c) to assert with scientists the importance necessity of proper model building, checking and reporting; (d) to insist that paragraphs discussing these points be included in publications as a matter of course.

Without these efforts, statistics risks losing it's credibility as a scientific discipline and as a consequence may also end up losing financial support.

71

Estimators For Partially Observed Markov Chains and Semi-Markov Processes

Ursula U. Müller, Anton Schick and Wolfgang Wefelmeyer

Fachbereich Mathematik und Informatik, Universität Bremen
Department of Mathematical Sciences, Binghamton University
Mathematisches Institut, Universität zu Köln

Abstract: Suppose we observe a semi-Markov process or a Markov renewal process at certain time points only. Which observation patterns allow us to identify the transition distribution of the embedded Markov chain or the conditional inter-jump time distribution? In case we can identify them, how can we estimate functionals of them from our observations? For smooth functionals, what is the best use we can make of the observations, at least asymptotically? We give an overview of possible approaches to these questions.

Keywords and phrases: Markov chain, semi-Markov process, Markov renewal process, partial observation, periodic skipping, random skipping, "skipping at random", empirical estimator, efficient estimator

71.1 Introduction

In this extended abstract, in order to simplify the exposition, we will concentrate mainly on Markov *chains* (with arbitrary state space), and on *nonparametric* models for them. We will not say much about *efficiency*. We will also restrict attention to estimating *linear* functionals of two successive realizations under the stationary law. This is not a serious restriction: The distribution of the chain is determined by the values of a sufficiently large class of such functionals.

In Section 71.2 we recall the case of fully observing the chain up to a fixed time. In the following sections we treat four different patterns of picking observations: periodic; random without knowing the clock of the chain; known random time points; and skip lengths that depend on the previous state of the chain.

Throughout, let $X_1, X_2, \ldots$ be realizations of a Markov chain with transition distribution $Q(x, dy)$. We assume that the chain is geometrically ergodic in L_2

and (for simplicity) strictly stationary. Let $h(X_1, X_2)$ be square-integrable. We consider the problem of estimating the expectation $Eh(X_1, X_2)$.

71.2 Full Observations

Suppose we observe $X_1, \ldots, X_{n+1}$. A natural estimator of $Eh(X_1, X_2)$ is the empirical estimator

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i, X_{i+1}).$$

It is efficient in the nonparametric model (i.e., if nothing is known about Q aside from appropriate ergodicity properties and, perhaps, certain smoothness assumptions); for different proofs see Penev (1991), Greenwood and Wefelmeyer (1995) and Bickel and Kwon (2004); for Markov step processes and semi-Markov processes see Greenwood and Wefelmeyer (1994, 1996).

71.3 Periodic Skipping

Suppose we observe only some of the realizations, in a deterministic pattern that repeats itself periodically, say with period m. Specifically, in the first period we observe at k times $1 \le j_1 < \cdots < j_k \le m$ and then at times $m + j_1, \ldots,$ $2m + j_1, \ldots,$ for $n + 1$ periods, say. (Here it is understood that we *know* how many realizations we skip. We will consider in Section 71.4 a pattern where we do not have this information.) The *skip lengths* are

$$s_1 = j_2 - j_1, \ldots, s_{k-1} = j_k - j_{k-1},\ s_k = m + j_1 - j_k.$$

a) In the simplest case, some of the skip lengths are 1. For example, let $m = 3$, $k = 2$, $j_1 = 1$, $j_2 = 2$. Then every third realization is missing. A simple estimator of $Eh(X_1, X_2)$ is the empirical estimator based on observed pairs of successive realizations of the chain,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_{3i-2}, X_{3i-1}).$$

This estimator is not efficient (unless the observations are independent). The information in the non-adjacent pairs (X_{3i-1}, X_{3i+1}) can be used as follows. Suppose the state space is real and $Q(x, dy)$ has density $q(x, y)$. Introduce

$$h_l(x, z) = E(h(X_2, X_3)|X_2 = x, X_4 = z) = \frac{\int q(x, y)q(y, z)h(x, y)\, dy}{\int q(x, y)q(y, z)\, dy}.$$

We have $Eh_l(X_2, X_4) = Eh(X_2, X_3)$. Plug in a (kernel) estimator $\hat{q}$ for the transition density q, based on the pairs (X_{3i-2}, X_{3i-1}), to obtain an estimator $\hat{h}_l(x, z)$. This gives rise to a new estimator of $Eh(X_1, X_2)$, namely

$$\frac{1}{n}\sum_{i=1}^{n} \hat{h}_l(X_{3i-1}, X_{3i+1}).$$

A third estimator is obtained from $h_r(x, z) = E(h(X_3, X_4)|X_2 = x, X_4 = z)$. The three estimators can be combined to improve on the first.

b) Suppose that none of the skip lengths is 1, but they have no common divisor. Then we can represent 1 as a linear combination of skip lengths. Suppose, for example, that $m = 5$, $k = 2$, $j_1 = 1$, $j_2 = 3$. Then the skip lengths are $s_1 = 2$, $s_2 = 3$, and, since $1 = 3 - 2$, we can write $Q = Q^{-2}Q^3$. To estimate the inverse of a transition distribution, decompose the state space into a finite number of sets and invert the corresponding empirical transition matrix.

c) If the skip lengths have a common divisor, Q is not identifiable. Suppose, for example, that $m = 2$, $k = 1$, $j_1 = 1$. Then we skip every second realization. The remaining observations allow us to estimate Q^2, but this does not identify the root Q uniquely. (In certain parametric and semiparametric models we can however still identify Q, for example if the chain follows a first-order linear autoregressive model.)

71.4 Random Skipping, Unknown Times

Suppose that, after an observation at time j, we make the next observation at time $j + s$ with probability a_s, but we do not observe the skip length s. Then our observations follow a Markov chain with transition distribution given by the mixture $R = \sum_{s\geq 1} a_s Q^s$. (This is a badly designed experiment. It is however close to observing a semi-Markov process at fixed times, in which case we also do not observe the number of jumps between successive observations of the process. For Markov jump processes see Bladt and Sørensen, 2005.) Suppose we *know* the probabilities a_s. (A geometric distribution might be plausible.) Then we can try to solve R for Q. For example, if $a_1 = p$ and $a_2 = 1 - p$, then $R = pQ + qQ^2$ with $q = 1 - p$, and

$$Q = \frac{1}{p}\Big(R - \frac{q}{p^2}R^2 + 2\Big(\frac{q}{p^2}\Big)^2 R^3 - \dots\Big).$$

.

71.5 Random Skipping

In the Markov chain setting, it is more realistic to observe also the skip lengths. Write $A(\{s\}) = a_s$ for their distribution. Then we observe pairs (S_i, X_i) that

form a Markov chain with transition distribution

$$R(x, ds, dy) = A(ds)Q^s(x, dy).$$

We can estimate A empirically. Estimation of $Eh(X_1, X_2)$ is similar to the case of periodic skipping in Section 71.3; it depends on the properties of the set of skip lengths with positive probabilities. In particular, if a_1 is positive, a simple estimator is the empirical estimator

$$\frac{1}{m} \sum_{S_i=1} h(X_i, X_{i+1})$$

with m the observed number of skip lengths $S_i = 1$.

71.6 "Skipping At Random"

In the previous section we have assumed (implicitly) that the skip lengths are independent of the Markov chain. It is however conceivable that the skip lengths depend on the previous state. Let $A(x, ds)$ denote the skip length distribution out of state x. Then we observe pairs (S_i, X_i) with transition distribution

$$R(x, ds, dy) = A(x, ds)Q^s(x, dy).$$

This factorization is analogous to the factorization $Q(x, dy)A(x, y, ds)$ of the transition distribution of a Markov renewal process; for efficient estimation in semiparametric models of the corresponding semi-Markov process see Greenwood *et al.* (2004). The name "skipping at random" is chosen because of the similarity with responses "missing at random" in regression models; for efficient semiparametric estimation see Müller *et al.* (2006). The case of Section 71.5, with A not depending on x, would correspond to "missing totally at random". Again, if $a_1(x) = A(x, \{1\})$ is positive with positive probability, a simple estimator of $Eh(X_1, X_2)$ can be based on the observed pairs of successive observations:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(S_i = 1)}{\hat{a}_1(X_i)} h(X_i, X_{i+1})$$

with $\hat{a}_1(x)$ a (kernel) estimator of the conditional probability $a_1(x) = A(x, \{1\})$.

References

1. Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: Some questions and an answer (with discussion), *Statistica Sinica*, **11**, 863–960.

2. Bladt, M. and Sørensen, M. (2005). Statistical inference for discretely observed Markov chain jump processes, *Journal of the Royal Statistical Society: Series B*, **67**, 395–410.

3. Greenwood, P. E., Müller, U. U. and Wefelmeyer, W. (2004). Efficient estimation for semiparametric semi-Markov processes, *Communications in Statistics Theory and Methods*, **33**, 419–435.

4. Greenwood, P. E. and Wefelmeyer, W. (1994). Nonparametric estimators for Markov step processes. *Stochastic Processes and their Applications*, **52**, 1–16.

5. Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains, *Annals of Statistics*, **23**, 132–143.

6. Greenwood, P. E. and Wefelmeyer, W. (1996). Empirical estimators for semi-Markov processes. *Mathematical Methods of Statistics*, **5**, 299–315.

7. Müller, U. U., Schick, A. and Wefelmeyer, W. (2006). Imputing responses that are not missing, In *Probability, Statistics and Modelling in Public Health* (Eds., Nikulin, N., Commenges, D. and Huber, C.), 350-363, Springer, New York.

8. Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains, *Journal of Statistical Planning and Inference*, **27**, 105–123.

72

New Weakest Link Distribution Family

Yuri Paramonov and Janis Andersons

Aviation Institute of Riga Technical University
Institute of Polymer Mechanics, University of Latvija

Abstract: A family of weakest-link models based on the assumption of a two-stage failure process is derived. In the first stage initiation (either instant or step by step) of some flaws in one or several links, and in the second stage fracture of the weakest link take place. The offered models sometimes more adequately describe the experimentally observed fiber strength scatter and the strength dependence on fiber length than the traditional models.

Keyword and phrases: Weakest-link, fiber tensile strength, cdf

72.1 Introduction

Power-Weibull (PW) model of distribution

$$F(x) = 1 - exp(-(l/l_0)^{\gamma}(x/\beta)^{\alpha}), \tag{72.1.1}$$

which was intensively studied in literature , while providing a good empirical fit to the strength data of specimens with different length l, lack the theoretical appeal of the weakest-link models. We derive a new family of weakest-link models based on the assumption of a two-stage failure process. For modelling purposes we consider a specimen as a chain of n elements (links) of length l_1. First, the process develops along the specimen and in K elements, $K \leq n$, flaws appear. In the second stage in the weakest element the accumulation of elementary damages takes place in crosswise direction up to specimen failure. We consider four different models: two versions of the first stage development in time (flaws in some elements appear instantly or gradually); two versions of the process development along the specimen (either in several crossections or only in one crossection).

72.2 Model of instant destruction

72.2.1 Model of instant destruction in the presence of several flaws

Let K, $0 \leq K \leq n$, be the number of elements in which flaws appear. K is a random variable. Let $Y_1, Y_2, ..., Y_K$ be strengths of these elements, X - strength of specimen. In this model we define

$$X = \begin{cases} \min(Y_1, Y_2, ..., Y_K) & if \quad 0 < K \leq n, \\ -\infty, & if \quad K = 0. \end{cases}$$

The mechanical stress is uniformly distributed along the fiber, therefore binomial distribution of the random variable K is expected. The corresponding probability mass function is $p_k = p^k(1-p)^{(n-k)}n!/k!(n-k)!$, where $p = F_0(x)$, $F_0(x)$ is the cumulative distribution function (cdf) of flaw initiation stress. Then the cdf of specimen strength, $F(x)$, is defined by the equation

$$F(x) = \sum_{k=0}^{n} p_k(1 - (1 - F_1(x))^k), \tag{72.2.2}$$

where $F_1(x)$ is the strength cdf of the fiber element with length l_1. If n is sufficiently large then the binomial distribution can be replaced by the Poisson distribution with $\lambda = np$. Note that this approximation of binomial distribution is unsatisfactory if n is small. For this reason we suggest a modified version of such an approximation

$$F(x) = \sum_{k=0}^{\infty} \frac{\lambda^k exp(-\lambda)}{k!}(1 - (1 - F_1(x))^{k+1}). \tag{72.2.3}$$

Eq. (72.2.3) need not be treated only as an approximation of eq. (72.2.2). It can be endowed with a specific interpretation. Namely, in eq. (72.2.2) the cdf $F_1(x)$ is the cdf of tensile strength of one element (link) of length l_1, and as such it would be expected to depend on l_1 which is rather inconvenient. Contrarily, eq. (72.2.3) can be interpreted as the cdf of tensile strength of a specimen in which during the first stage of the failure process K weak cross sections (WCS) are initiated, $K = 0, 1, 2, ,$ in a corresponding Poisson process. Then the function $F_1(x)$ is interpreted as the cdf of the tensile strength of a WCS. Importantly, here $F_1(x)$ characterizes failure of the WCS, and therefore should depend neither on the length of the link, l_1 , nor on the total length of specimen. Eqs. (72.2.2) and (72.2.3) present a family of two subfamilies of distributions defined by the choice of either binomial or Poisson distribution and by the choice of a pair of functions ($F_0(x)$,$F_1(x)$).

72.2.2 Model of instant destruction in the presence of at least one flaw

It can be assumed also that the distribution of flaws is uniform only at the beginning of the process but then the destruction process develops only in one element. We make an additional assumption that the strength of this critical element, Y, is random variable the cdf of which does not depend on the length of specimen. Let I_j, $j = 1, 2, ..., n$, be random variables equal to 1 if there is a flaw in jth element and equal to 0 if there is no flaw in this element. According to these assumptions, the strength of specimen

$$X = \begin{cases} Y, & if \quad max(I_1, I_2, ..., I_n) = 1, \\ \infty, & if \quad max(I_1, I_2, ..., I_n) = 0. \end{cases}$$

Then

$$F(x) = (1 - (1 - F_0(x))^n)F_1(x), \tag{72.2.4}$$

where $F_0(x)$ and $F_1(x)$ are the same as in previous subsection.

72.3 Model of step by step accumulation of flaws

72.3.1 Model of destruction in the presence of several flaws

We consider the process of accumulation of flaws as an inhomogeneous finite Markov chain (MC) with finite state space $I = \{i_1, i_2, ..., i_n, i_{n+1}, i_{n+2}\}$. We say that the MC is in state i_k if there are $(k-1)$ flaws, $k = 1, ..., n+1$. State i_{n+2} is an absorbing state corresponding to destruction of specimen. Usually we suppose that the Markov chain starts in state i_1 but in general case the initial distribution is represented as a row vector π given by $\pi = (\pi_1, \pi_1, ...\pi_{n+2})$. We further assume that the loading (i.e. the process of nominal stress increase in the specimen cross section) is described by an ascending (up to infinity) sequence $\{x_1, x_2, ...\}$ and the transition probabilities p_{ij} of the transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1(n+1)} & p_{1(n+2)} \\ 0 & p_{21} & p_{23} & \dots & p_{2(n+1)} & p_{2(n+2)} \\ 0 & 0 & p_{31} & \dots & p_{3(n+1)} & p_{3(n+2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & p_{(n+1)(n+1)} & p_{(n+1)(n+2)} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

at the m-step are functions of x_m. Let the sequance $\{x_m\}$ be fixed. Then we denote them as function of m. The probability that in the some element a flaw appears at the stress x_m under the condition that it has not appeared at the stress $x_{(m-1)}$

$$b(m) = (F_0(x_m) - F_0(x_{(m-1)}))/(1 - F_0(x_{(m-1)})).$$

Consider the case of s flaws present. The probability that r new flaws appear, $0 \le r \le k = n - s$, and the total number of flaws equals to $t = s + r$

$$\tilde{p}_{st}(m) = (b(m))^r (1 - b(m))^{(k-r)} k!/r!(k-r)!.$$

Conditional probability of some element destruction at the nominal stress x_m

$$q(m) = (F_1(x_m) - F_1(x_{(m-1)}))/(1 - F_1(x_{(m-1)})).$$

Corresponding probability that destruction of no element takes place when there are flaws in t elements

$$u_t(m) = ((1 - q(m))^t.$$

The probability of coincidence of these events, which we consider as independent, is the probability of transition from state $i = (s + 1)$ to state $j = i + r$

$$p_{ij}(m) = \tilde{p}_{(i-1)(j-1)}(m) u_{j-1}(m),$$

where $i \le j \le (n + 1)$. Conditional destruction probability at state i

$$p_{i(n+2)}(m) = 1 - \sum_{j=i}^{n+1} p_{ij}(m).$$

Of course, $p_{ij}(m) = 0$ if $j < i$ and $p_{(n+2)(n+2)}(m) = 1$.

72.3.2 Model of destruction in the presence of at least one flaw

In this case in MC there are only three states: there are no flaws (state number $i = 1$), there is at least one flaw (state number $i = 2$), absorbing state (state number $i = 3$). Corresponding transition probabilities are: $p_{11}(m) = (1 - b(m))^n$, $p_{12}(m) = 1 - p_{11}(m)$, $p_{13} = 0$, $p_{21} = 0$, $p_{22}(m) = 1 - q(m)$, $p_{23}(m) = q(m)$, $p_{31} = p_{32} = 0$, $p_{33} = 1$.

For both models considered in subsections 72.3.1 and 72.3.2 the cdf of specimen strength (defined on the sequence $\{x_m\}$) is defined actually by the step number to absorption

$$F(x_m) = \pi(\prod_{j=1}^{m} P(j))u, \tag{72.3.5}$$

where $P(j)$ is transition matrix for step number j, column vector $u = (0...01)'$ where only the last component is equal to 1 but all the others are equal to 0.

72.4 Model parameter estimation. Comparison of models

Considered models were used for processing of dataset described in Andersons *et al.* (2002) and (2005) for glass fiber and flax fiber tests respectively. It was assumed that

$$F_0(x) = F_1(x) = 1 - exp(-exp((x - \vartheta_0)/\vartheta_1)),$$

where $x = log(\sigma)$, σ is strength (MPa). In this case $F(x)$ depends on l_1 and on location and scale parameters, ϑ_0 and ϑ_1, estimates of which (for fixed l_1) can be found using regression analysis of order statistics. The maximum likelihood method can be used here also but it is excessively labor-consuming. Our purpose is only comparison of the models, and we have limited ourselves by the use of regression analysis.
Let X_{ij} be jth order statistic in a sample corresponding to specimen length $L = L_i$, $E(X_{ij})$ is expected value of X_{ij}, $E(\overset{0}{X}_{ij})$ is the same but for $\vartheta_0 = 0$ and $\vartheta_1 = 1$, $\hat{\vartheta}_0$ and $\hat{\vartheta}_1$ are estimates of ϑ_0 and ϑ_1, $\hat{x}_{ij} = \hat{\vartheta}_0 + \hat{\vartheta}_1 E(\overset{0}{X}_{ij})$ is estimate of $E(X_{ij})$, $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n$, $\hat{x}_i = \sum_{j=1}^{n_i} \hat{x}_{ij}/n$, n_i is number of specimens with $L = L_i$, $i = 1, 2, ..., k_L$, where k_L is number of different L. The values of l_1, for which we have minimum of values $Q_1 = (\sum_{i=1}^{k_L}(\bar{x}_i - \hat{x}_i)^2 / \sum_{i=1}^{k_L}(\bar{x}_i - \bar{x})^2)^{1/2}$ or $Q_2 = (\sum_{i=1}^{k_L}\sum_{j=1}^{n_i}(x_{ij} - \hat{x}_{ij})^2 / \sum_{i=1}^{k_L}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2)^{1/2}$, where $\bar{x} = \sum_{i=1}^{k_L} \bar{x}_i/k_L$,, can be used as estimates of parameter l_1. We use the corresponding values of Q_1 and Q_2 for model comparison (Remark. If only two parameters, ϑ_0 and ϑ_1, are unknown, then the value Q_2 can be used for goodness-of-fit test of cdf type hypothesis testing).

It is appears that the best fit for the four samples of glass fiber strengths (78,74, 50 and 60 observations, Andersons *et al.* (2002), with l equal to 10, 20, 40 and 80 mm correspondingly) was provided by eq. (72.1.1) for criterion Q_2 and by eq. (72.2.3) for criterion Q_1 . The best fit (Fig.1) for the three samples of flax fiber strength (90, 70 and 58 observations, Andersons *et al.* (2005), with L equal to 5,10 and 20 mm correspondingly) was provided by eq. (72.2.4) for criterion Q_2 ($Q_2 = 0.1718$ and by eq. (72.2.3) for criterion Q_1 ($Q_1 = 0.1261$). The models with the use of MC theory give approximately the same results.
Evidently, we have random conclusions because we have random test data. But it seems that all four considered models deserve to be studied much more

thoroughly, using much more test data.

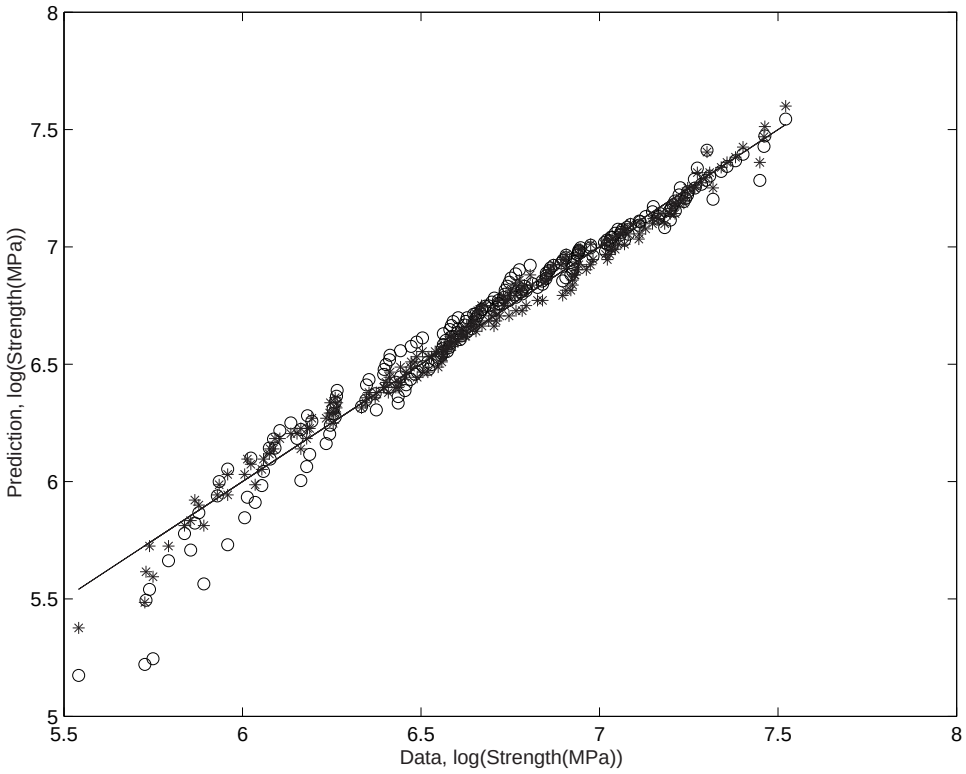


Figure 1: Comparison: x_{ij} (Andersons *et al.* (2005))and $\hat{x}_{ij}$, calculated using eqs (72.1.1) and (72.2.4) denoted by (o) and (*) correspondingly.

References

1. Andersons, J., Joffe, R., Hojo, M. and Ochiai, S. (2002). Glass fibre strength distribution determined by common experimental methods, *Composites Science and Technology* , **62**, 131–145.

2. Andersons, J., Sparnins, E., Joffe, R.and Wallstrom L. (2005). Strength distribution of elementary flax fibers, *Composites Science and Technology*, **65**, 693–702.

73

Bayes - Fiducial Approach For Quantile Estimation And Specified Life Nomination

Yuri Paramonov

Aviation Institute of Riga Technical University

Abstract: An optimality of specified life, t_{SL}, nomination decision is discussed. Two criteria are considered: maximum of expectation of t_{SL} under limitation of failure probability and maximum of operation time expectation, when the failure is estimated by some negative value. Numerical examples are given.

Keyword and phrases: quantile estimation, specified life, Bayes-fiducial approach

73.1 Introduction

In this paper we consider the case when some system should be discarded from service if its service life exceeds some service time, t_{SL}, to which we'll refer as the *specified life (SL)*. It should be done in order to ensure the reliability of the system or in order to have the maximum of some profit when the failure of the system before t_{SL} is connected with some big loss. SL should be specified on the base of processing of some timetest data. The exposition is as follows. Application of p-bound to solution of the problem of the SL nomination under condition of failure probability limitation is discussed in section 2. Economics approach is considered in section 3.

73.2 Specified life nomination under condition of failure probability limitation

P-bound is special case of p-set function, definition of which is given in Paramonov and Paramonova (1998), where p-set function is used for inspection

programme development.
The p-bound for random variable is defined in the following way.

P-bound definition. Let Z be a random variable (rv) and $X = (X_1, X_2, ..., X_n)$ be a random vector. We suppose that it is known the class $\{P_\theta, \theta \in \Omega\}$ to which the probability distribution of the random vector $W = (Z, X)$ is assumed to belong. Of the parameter θ, which labels the distribution, it is assumed known only that it lies in a certain set Ω, the parameter space.

Function $\tau(x)$ is called a p-bound for the rv Z if $\sup_\theta P(Z < \tau(X)) = p$, $\theta \in \Omega$. It is called a parameter-free (p.f.) p-bound for r.v. Z if $P(Z < \tau(X)) = p$ for all $\theta \in \Omega$.

So really, it is a p-quantile of cdf $F_Z(z)$ estimate. But it is very specific estimate: expectation value $E(F_Z(\tau(X))) = p$. P.f. p-bound has close connection with prediction interval (see Paramonov and Paramonova (1998)) but in the problem considered we have very specific loss function. We are interested to get the maximum of expectation value of p-bound under condition $E(F_Z(\tau(X))) = p$.

Later on the value x, observation of the random vector X, would be interpreted as result of some test, some sample; random value Z would be interpreted as lifetime of the considered system in service; the value $\tau(x)$ as nominated specified life; the probability $P(Z < \tau(X))$ as the probability of failure; condition $P(Z < \tau(X)) = p$ as condition to provide the required reliability equal to $(1 - p)$.

Let Z and every component of vector X have the following structure $Z = \theta_0 + \theta_1 \overset{0}{Z}$, $X_i = \theta_0 + \theta_1 \overset{0}{X}_i$, where θ_0, θ_1 are unknown location and scale parameters, cdf of r.v. $\overset{0}{Z}$ and $\overset{0}{X}_i$, $i = 1, ..., n$, are known.

Let $\tau(x))$ be linear function of order statistics $\tau(x) = ax_{(1,...,n)}$, where $a = (a_1, ..., a_n)$ is row vector, $x_{(1,...,n)} = (x_{(1)}, ..., x_{(n)})^T$ is column vector of order statistics $X_{(1,...,n)} = (X_{(1)}, ..., X_{(n)})^T$ (here the transpose (of a vector or of a matrix) is denoted by a capital superscript T). If $a\xi = 1$, where $\xi = (1, ..., 1)^T$ is column vector of n units, then $\tau(x)$ is p.f. p-bound for r.v. Z for some p, because in this case $\tau(X)$ has the following structure: $\tau(X) = \theta_0 + \theta_1 \overset{0}{\tau}$,

where $\overset{0}{\tau} = \tau(\overset{0}{X}) = a \overset{0}{X}_{(1,...,n)}$, $\overset{0}{X}_{(1,...,n)}$ has the same cdf as $X_{(1,...,n)}$ but $\theta_0 = 0$, $\theta_1 = 1$. Let $\overset{0}{\nu}$ is the expectation value of $\overset{0}{\tau}$. Using the theorem 1.f.1(II) in Rao (1968) we can get the vector $a = a(\overset{0}{\nu})$ corresponding to the minimum of variance of $\overset{0}{\tau}$ at the fixed $\overset{0}{\nu}$:

$$a(\overset{0}{\nu})^T = \overset{0}{D}^{-1} BS^{-1}W,$$

where $\overset{0}{D}$ is a matrix of covariance of order statistics $\overset{0}{X}_{(i)}$, $i = 1, ..., n$; matrix

$B = (\overset{0}{\mu}, \xi)$, $\overset{0}{\mu} = (\overset{0}{\mu}_{(1)}, ..., \overset{0}{\mu}_{(n)})^T$ is column vector of expectation values of order statistics for $\theta_0 = 0$, $\theta_1 = 1$, matrix $S = B^T \overset{0}{D}^{-1} B$, column vector $W = (\overset{0}{\nu}, 1)^T$.

Then $P(Z < \tau(X)) = P(\overset{0}{Z} - \tau(\overset{0}{X}) \leq 0) = P(U \leq 0) = p(\overset{0}{\nu})$.
If random variable U has normal distribution then for fixed $P(U \leq 0)$, provided $P(U \leq 0) \leq 0.5$, we have maximum of $\overset{0}{\nu}$ and expectation value of $\tau(X)$. Evidently, we have got only approximate optimum vector a if rv U has approximate normal distribution, but for chosen a the true value of probability $P(U \leq 0)$ can be calculated using, for example, Monte Carlo method.

Let us consider the following numerical example. Suppose that simultaneously fatigue tests of 6 airframes of the same type of aircraft have been made but only until 4-th fatigue failure. So we know only 4 first minimal fatigue lives: $(t_{(1)}, ..., t_{(4)}) = (59971; 72600; 77630; 80863)$ and correspondingly $x = (x_{(1)}, ..., x_{(4)}) = (11.002; 11.193; 11.260, 11.3005)$, where $x_{(i)} = \ln\ t_{(i)}$, $i = 1, ...4,$. There are $m = 100$ aircraft in operation. There is requirement, that the probability of at least one fatigue failure up to specified life should not exceed $p = 0.05$.

Let us suppose, that lognormal distribution functions can be used for fatigue life data processing. We'll use the logarithm scale and correspondingly consider normal distribution. Expectation values and covariance matrix for first fourth order statistics of 6 observations from standard normal distribution can be found in special tabels or can be calculated using, for example, SAS or MATLAB . Using Monte Carlo (MC) method we can find that $\overset{0}{\nu}_p = -7.0$ and vector $a = [3.8883; 1.5865; 0.3789; -4.8537]$ for $p = 0.05$. In logarithm scale $\tau = 9.9539$. In natural scale $t_{SL} = 21,035$.

Now we consider decision based on the use of sufficient statistics and Bayes-fiducial approach. Let θ_1 is known. The random variable $\widehat{\theta}_t = \tau(x) = \widehat{\theta}_0 + t\theta_1$ is unbiased estimate of its own expectation (some quantile θ_t). If estimate $\widehat{\theta}_t$ is function of sufficient statistics then, as it is well known, we have minimum risk if correspondent loss-function is convex. In considered problem the function $F_Z(z)$ can be considered as a loss-function, which, for example, for normal distribution is convex (and increasing one) if $z < c = 0.5$. Then expectation $E^X\{F_Z(\widehat{\theta}_t)\} = P(Z < \tau(X))$ is the risk function. Let us use such parameter estimate $\widehat{\theta}_t$ of θ_t, which gives minimum risk. Then if we set the mimimum risk equal to p, we get the maximum value of $E\{\tau(X)\}$ corresponding to this p if p is small enough (and probability $P(\tau(X) < c)$ is high enough, where c is such that $F_Z(z)$ is convex if $z < c$).

The straight way to get $\tau(x)$ as function of sufficient statistics is the use of fiducial distribution (see Rao (1968)). Let $\tau(x, p_0)$ be the solution of the equation

$$E^{\widetilde{\theta}}\{F_Z((\tau-\widetilde{\theta}_0)/\theta_1)=p_0,$$

where $\widetilde{\theta}=(\widetilde{\theta}_0,\theta_1)$, r.v. $\widetilde{\theta}_0$ have fiducial distribution.

Let $x_i=\theta_0+\theta_1 \overset{0}{x}_i$, $i=1,...,n$, then it is easy to show that there is such p_0 that τ will be p.f.p-bound for Z and it is function of sufficient statistic.
Similar approach can be used if θ_1 is not known also. Fiducial density for the location and scale parameters in this case is defined in the following way (see Paramonov (1992)):

$$f_{\widetilde{\theta}_0,\widetilde{\theta}_1}(s_0,s_1;x)=g(s_0,s_1;x)/\int\int g(s_0,s_1;x)ds_0ds_1,$$

where $g(s_0,s_1;x)=f((x_1-s_0)/s_1)...f((x_n-s_0)/s_1)/s_1^{(n+1)}$, $-\infty\le s_0\le\infty$, $0\le s_1\le\infty$.

73.3 Bayes-fiducial method in framework of economics approach

We suppose that the income of aircraft successful service during time t is equal to t, but in case of failure we suppose to have a loss, which is equal to some value $w=-b$, where b is a very large positive number. Then income of one aircraft service, r.v. U, is defined by formula

$$U=\begin{cases} t_{SL}, & if \quad T>t_{SL},\\ T-b, & if \quad T\le t_{SL},\end{cases}$$

where T is r.v., fatigue life of SSI, t_{SL} is some SL. Let $F_T(t,\theta)$ be c.d.f. of r.v. T. Then u, expectation value of r.v. U, as function of t_{SL}, θ, b is defined by formula

$$u(t_{SL},\theta,b)=\int_0^{t_{SL}}(t-b)dF_T(t,\theta)+t_{SL}(1-F_T(t,\theta)).$$

Maximum of $u(t_{SL},\theta,b)$ is reached at optimum value of SL, t^*_{SL}, which is the root of the equation

$$bf_T(t,\theta)/(1-F_T(t,\theta))=1.$$

Let us consider the normal distribution of $X=lnT$ with c.d.f. $F_X(x)=\Phi((x-\theta_0)/\theta_1)$, where $\Phi(.)$ is standard normal c.d.f.. Then for the known t^*_{SL} corresponding θ_0 is defined by formula

$$\theta_0=t^*_{SL}-\theta_1\lambda^{-1}(t^*_{SL}\theta_1/b),$$

where $\lambda(z)$ is failure rate function for standard normal distribution, $\lambda^{-1}(.)$ is inverse function. Now t^*_{SL} can be defined as the corresponding inverse function for wich we use the following notation :

$$t^*_{SL} = S^*_L(\theta_0, \theta_1, b). \tag{73.3.1}$$

For $b = 346000$, $\theta_1 = 0.346$ and $\theta_0 = 9.948$ maximum value of u corresponds to $t^*_{SL} = 7936$. It is interesting to note that this value corresponds to the probability failure 0.0026. This can be interpreted in the following way. Failure of 2.6 aircraft (in flight) from the park of 1000 aircraft can be considered as equivalent to the loss of 346000 hours of service time or loss of $346000/7936 = 43.6$ aircraft (on the ground) of the same types (the price of an aircraft is considered to be equal to $t_{SL} = 7936$). In other words, failure of one aircraft (in flight) is equivalent to the loss of $43.6/2.6 \approx 16$ aircraft of the same type (on the ground).

But we do not know the parameters and should estimate them using fatigue test data. Usually maximum likelihood estimate is considered as most appropriate. We show here that for the problem considered the Bayes-fiducial approach proposed (see Paramonov, 1992) is much more appropriate. In accordance with Bayes approach the parameter θ is considered as some rv. For the case of airframe it can be interpreted in the following way. Design stress analysis of an airframe should be made in accordance with some requirements (FAR, ...). These requirements in fact define only some mean value of θ_0, but of course, in every case there are some "occasional mistakes" and we have some specific (random) value of θ_0 for every aircraft type. And then there is a scatter of rv X at this random θ_0. The parameter θ_1 is function of technology level, and if one is not changed, then the parameter θ_1 is not changed also.

So suppose that θ_1 is known constant but θ_0 is random variable which we denote by $\tilde{\theta}_0$. Denote by $\pi(\theta_0)$ a priori distribution density of $\tilde{\theta}_0$, then c.d.f. of new r.v. $\tilde{X}$ will be

$$\tilde{F}_{\tilde{X}}(x) = \int_{-\infty}^{\infty} F_X((x - \theta_0)/\theta_1)\pi(\theta_0)d\theta_0.$$

It is easy to show that if θ_1 is constant, rv $\tilde{\theta}_0$ has normal distribution with mean τ_0 and standard deviation τ_1, then distribution of $\tilde{X}$ will be again normal with mean τ_0 and standard deviation $((\tau_1)^2 + (\theta_1)^2)^{1/2}$. In this case optimal SL, t_{SL}, again will be defined by the same eq. (73.3.1) but θ_0 should be replaced by τ_0 , θ_1 should be replaced by $((\tau_1)^2 + (\theta_1)^2)^{1/2}$.

In fact we do not know a priori distribution of $\tilde{\theta}_0$, but we have sample from $F_X(x)$ distribution. For this case the BF approach is offered. Instead of a

priori distribution of $\tilde{\theta}_0$, we offer to use the already mentioned fiducial distribution. The main ideas of this approach in case of two unknown parameters are described, for instance, in Paramonov (1992). In considered case, when θ_1 is known, it is enough to say that fiducial distribution of $\tilde{\theta}_0$ is normal again with mean $\bar{x}$ and standard deviation $\theta_1/n^{1/2}$. Then for the purpose of calculation of posterior t_{SL} we again can use the same eq. (73.3.1), but θ_0 should be replaced by $\bar{x}$ and θ_1 should be replaced by $\tilde{\theta}_1 = \theta_1(1+1/n)^{1/2}$.

Let us make comparison of BF approach with direct use of maximum likelihood (ML) estimates instead of θ. Let $t_{SL}(x)$ is some function of observation vector x and random variable U_X is defined by formula

$$U_X = u(t_{SL}(X), \theta, b).$$

For BF approach

$$t_{SL}(x) = S_L^*(\bar{x}, \tilde{\theta}_1, b).$$

If we use ML estimate of θ_0 then

$$t_{SL}(x) = S_L^*(\bar{x}, \theta_1, b).$$

By the use of Monte Carlo method for $\theta = (\theta_0, \theta_1) = (9.948, 0.346), b = 346,000$ for the sample size $n = 1, 2, 4, 100$ for BF approach we have got following expectation value of r.v. U_X, $E(U_X)$: 2310 4122 5571 6904. If we use ML estimate of θ_0 then the corresponding values of $E(U_X)$ are equal to -8624 809 4422 6935. We see that for small n the expectation value of rv U_X is much more for BF method than for ML method. The value of $E(U_X)$ is negative for the ML method and $n = 1$ because if t_{SL} more than t_{SL}^* then U decreases very drasticly. Standard deviation of $\bar{X}$ is equal to $\theta_1/n^{1/2}$ and for $n = 100$ it is very small. In this case we have nearly the same value of $E(U_X)$ as for the known θ_0 for both BF and ML methods.

References

1. Paramonov, Yu.M. and Paramonova, A.Yu. (1987). Inspection Program Development by the Use of Approval Test Results,*Int.J.of Reliability Quality and Safety Enginiering*, **62**, 301–308.

2. Rao, C.Radhakrishna (1968)*Linear Statistical Inference and its Applications*, John Wiley & Sons Inc.,New York-London-Sidney; Nauka, Moscow (Russian).

3. Paramonov,Yu.M. (1992) *Mathematical Statistics Methods for Estimation and Ensuring of the Reliability of Airframe*, Riga Aviation Institute, Riga (Russian).

74

Analyzing Non-Proportional Hazards

Aris Perperoglou and Hans C. van Houwelingen

Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands

Abstract: Several modelling techniques have been proposed for non proportional hazards. In this work we consider different models which can be classified into two wide categories: models with time varying effects of the covariates and frailty models. We present these different extensions of non-proportional hazards models on an application of 2433 breast cancer patients with a long follow up. We comment on the differences and similarities among the models and evaluate their performance using survival and hazard plots, Brier scores and pseudo-observations.

Keywords and phrases: reduced rank , relaxed Burr models, Brier Scores

74.1 Introduction

When analyzing survival data the Cox proportional hazards model is considered the 'null' model. However, in studies with long follow up of the patients the assumption of proportionality is often violated and alternative modelling strategies have to be considered. To present some of these approaches we use data from the breast cancer registry of IASO Woman's hospital in Athens, Greece. We consider a data set with 2433 patients operated for breast cancer with a maximum follow-up of 21 years.

We will analyze the data using Cox models with time varying effects, reduced rank models, the gamma frailty (Burr) model and the relaxed Burr model. We will shortly present these approaches and discuss their properties. We will try to highlight the similarities and pinpoint their differences when applied to the data.

To evaluate the different modelling strategies, we will compare survival and hazard plots computed from the models and discuss the use of Brier scores combined with pseudo-observations.

The data come from the registry of breast cancer patients of IASO Woman's hospital in Athens, Greece. From 1981 up to 2002, 2433 women with operable breast cancer were treated in the department of breast oncology of the institution. More information about the data can be found in Perperoglou et al (b,2006).

74.2 Time varying effects and reduced rank models

Perperoglou et al. (a,2006) introduced reduced rank regression models in survival analysis. A reduced rank model is a Cox model with time dependent effects of the covariates written as:

$$h(t|X) = h_0(t)\exp(X\Theta F') \tag{74.2.1}$$

with X a row-vector of p covariates, F a vector of q time functions, and Θ is a $p \times q$ matrix of estimable coefficients. Written in this way, matrix Θ can be factorized in several different ways, as the product $\Theta = B\Gamma'$, where B is a $p \times r$ and Γ a $q \times r$ matrix, and r is the rank of the reduced rank model. The maximum rank of the model can be $r = \min(p, q)$ resulting in the very flexible full rank model, while when the rank is smaller the model is more rigid since the number of parameters used to model the time varying effects is smaller.

74.3 The Burr and relaxed Burr model

The gamma frailty (Burr) model is very often used to describe individual heterogeneity or in general to account for the possible lack of fit. However, the assumption that frailties are constant might be restrictive and against biological reasoning. To account for time dependent frailties Perperoglou et al (c,2006) introduced the relaxed Burr model as an alternative more flexible model. Then relaxed Burr model is defined as

$$h(t|X) = \frac{h_0(t)\exp(X\beta)}{1 + F(t|\delta)\exp(X\beta)} \tag{74.3.2}$$

where $F(t|\delta)$ can be any continuous non-negative function starting at $F(0|\delta)=0$. We will always use a linear model $F(t|\delta) = f(t)\delta$ where $f(t)$ can be a simple time function multiplied by an unknown but estimable coefficient δ. This generalization of the Burr model allows for more flexible forms of hazard functions, depending on the functional form of $F(t|\delta)$.

74.4 Use of pseudo-observations and Brier scores

A Brier score measures average discrepancies between the true disease outcome and the predictive values from the model. To avoid the complications of censor-

ing weights are introduced when a case is censored at the time t where the scores are evaluated. We propose to compute Brier scores on pseudo-observations (Andersen 2003). For that purpose define the Brier *pseudo-observation* score:

$$Bp.s(t) = \frac{1}{n}\sum_{i=1}^{n}\{[P.O(t) - \hat{S}(t|X_i)]^2\} \tag{74.4.3}$$

where $P.O(t)$ are the pseudo-observations evaluated at time t. given by $P.obs_i(t) = nK\hat{M}(t) - (n-1)K\hat{M_{-i}}(t)$ whith KM the Kaplan-Meier survival estimate at time t and $\text{KM}_{-i}(t)$ the estimate at the same time, leaving individual i out.

74.5 Model comparison

The reduced rank model was fitted to the data using second degree B-splines with 3 interior knots as time functions. All possible reduced rank models were fitted, from the very rigid rank=1 model to the very flexible full rank model. The rank=2 model has 22 free parameters and was chosen to analyze the data. When dealing with time varying effects of the covariates, expressing relationships with a single number is more complicated, as well as building prognostic scores. However, a plot of the covariate effects through time can be very useful and offer an insight on the nature of the data. In figure (74.1) the effects of covariates under a rank=2 model are presented.

We fitted a gamma frailty model to the data and the relaxed Burr model using cubic B-splines as time functions. The estimates of the coefficients and their standard errors were very similar under the two models (data not shown). However, the relaxed Burr model is more flexible and shows that the frailties are not constant but change through time. To show this consider two hypothetical group of patients that emerge from the data, a group of 50 year old individuals, with tumor size of 15mm and 3 positive lymph nodes with the difference that in the first group the tumor grading was I, while in the second the tumor was of grade 3. Under a proportional hazards model, these two different group of patients will have parallel hazards regardless of time. This can be seen in figure (74.2) where the lower solid line presents the hazard under a Cox proportional hazards model, for patients in the first group, and the upper solid line presents the hazard for patients in the second group. In the same graph the dashed lines present the hazards of the two different groups under the Burr model, and the pointed lines are the hazards coming from a relaxed Burr model. As it is expected the Burr model shows hazards that converge after approximately 10 years. The relaxed Burr model gives almost identical hazards to the Burr model for the first 5 years, but from that point on the hazards start to diverge and come closer to the ones given by the simple Cox model. That means that after the elapse of some time, the group of people that remain do not have constant

frailties anymore, instead that frailties change and the model comes close to proportional hazards.

To compare the models with respect to survival we created several different group of patients, running from good to worse prognosis, all of which emerge from the data. In most of the cases the differences among the models were hardly visible. Here we present a group of patients with aged 66 years old with tumor size 30mm, 7 positive lymph nodes and grade 3, and present their survival functions in Figure 74.3. As it can be seen, up to the fifth year of follow-up the models give very similar estimates while from the sixth year and on, the reduced rank model is giving larger probabilities of survival than the rest as it is expected since the effects of the covariates are weakened under after some time. The differences among the Burr model and relaxed Burr model are hardly visible for up to 6 years in all patient groups, while from the seventh year and on are very small.

The Brier pseudo-observation scores at different time points are presented in table (74.1). The scores show that the reduced rank model is the best up to the fifth year but from that point on that the relaxed Burr model is the preferred one.

References

1. A. Perperoglou, S. le Cessie, and H. C. van Houwelingen, (2006). Reduced Rank hazard regression for modelling non-proportional hazards, *Statistics in Medicine*, in press, DOI: 10.1002/sim.2360

2. Perperoglou A, le Cessie S., van Houwelingen H.C. (2006). A fast routine for fitting Cox models with time varying effects of the covariates, *Computer methods and programs in biomedicine*, **81**, 154–161.

3. A. Perperoglou, H. C. van Houwelingen, and R. Henderson, (2006) A relaxation of the gamma frailty (Burr) model *Submitted*

4. P. K. Anderesen, and J. P. Klein, (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models, *Biometrica*, 15-27.

Table 74.1: Brier score and R^2 for different models estimated on pseudo-observations. Time (t) in months.

	KM	PH	Burr	RBurr	Rank=2
$t = 24$	0.0598	0.0563	0.0558	0.0561	0.0554
$t = 36$	0.1102	0.1013	0.1004	0.1006	0.1000
$t = 48$	0.1472	0.1359	0.1352	0.1352	0.1350
$t = 60$	0.1809	0.1657	0.1648	0.1646	0.1649
$t = 72$	0.2215	0.2041	0.2032	0.2027	0.2041
$t = 84$	0.2690	0.2478	0.2471	0.2463	0.2489
$t = 96$	0.3032	0.2797	0.2788	0.2780	0.2814
$t = 120$	0.3858	0.3602	0.3592	0.3581	0.3633

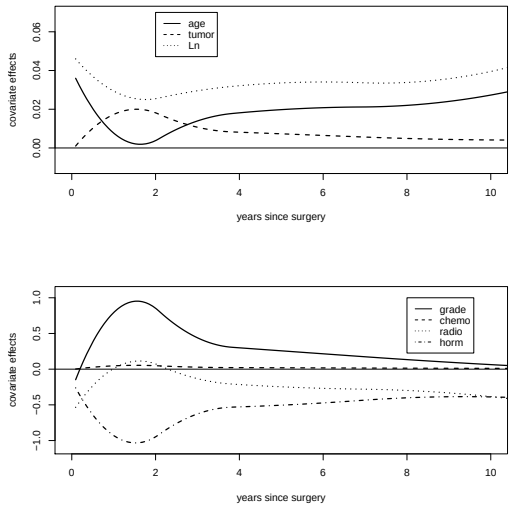


Figure 74.1: Effects of covariates through time under the rank=2 model.

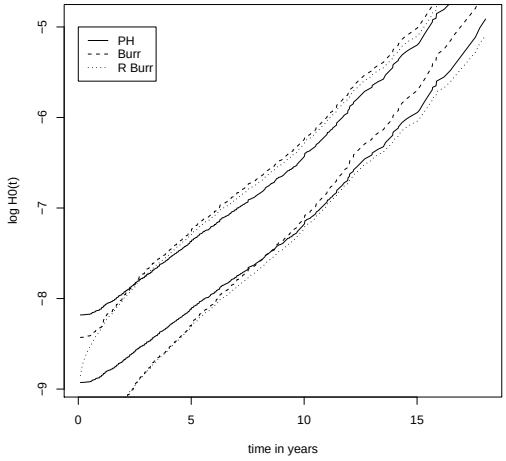


Figure 74.2: Plot of log hazards for two different patient groups.

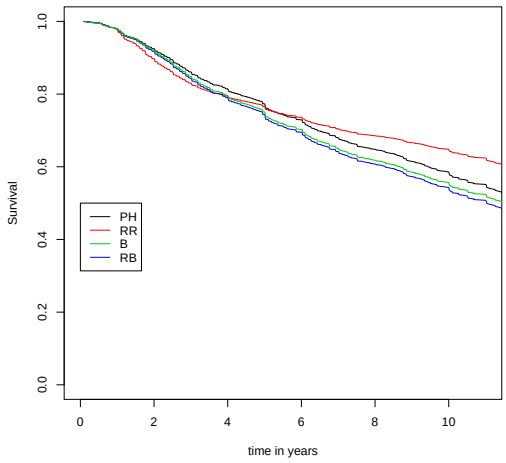


Figure 74.3: Plot of survival functions for four different groups under four different models; PH:proportional hazards, RR: rank=2 model, B: Burr and RB: relaxed Burr model.

75

Generalized Birth and Death Processes as Degradation Models

Vladimir Rykov

Institute for Information Transmission Problems RAS, Moscow

Abstract:[1] To model degradation processes in technical and biological objects generalized birth and death processes are introduced and studied.
Keywords and phrases: Birth and death processes generalization, Degradation models.

75.1 Introduction and Motivation

Traditional studies of technical and biological objects reliability mainly deals with their reliability function and steady state probabilities for renewable systems. Nevertheless, because there are no infinitely long leaving objects and any repair is possible only from the state of partial failure, the modelling of degradation process during a life period of an object is a mostly interesting topic. From the mathematical point of view the degradation during object's life period can be described by the Birth and Death (B & D) type process with absorbing state. For this process the conditional state probability distribution given object's life period is a mostly interesting characteristic.

During last years an intensive attention to the aging and degradation models for technical and biological objects has been attracted. The organization of special scientific conferences devoted to this topics testifies it. The aging and degradation models suppose the study of the systems with gradual failures for which multi-state reliability models were elaborated (for the history and bibliography see, for example, Lisniansky and Levitin (2003)). In some of our previous papers (see Rykov and others (2004) and the bibliography therein) the model of complex hierarchical system was proposed and the methods for its steady state and time dependent characteristics investigation was done. Con-

[1]The paper was partially supported by the RFFI Grant No. 04-07-90115.

trollable fault tolerance reliability systems were considered in Rykov Efrosinin (2004) and Rykov Buldaeva (2004). In the present paper a generalized B & D process as a model for degradation and aging processes of technical and biological objects is proposed. Conditional state probabilities given object's life period and their limiting values when $t \to \infty$ are calculated. The variation of the model parameters allows to consider various problems of aging and degradation control.

75.2 Generalized Birth & Death Process

Most of up-to-date complex technical systems also as biological objects with sufficiently high organization during their life period pass over different states of evolution and existence. From reliability point of view these states can be divided into three groups: the states of normal functioning, the dangerous (degradation) states and the failure states, that can be joined into the sets N, D, and F respectively.

In the simplest case if the nature of the degradation process allows to completely order the states to admit the transition possibilities only to neighboring states it can be modelled by the process of B & D type.

75.2.1 Definition. Basic equalities

Suppose that the states of the object are completely ordered, its transitions only into neighboring states are possible, and their intensities depend on the time spend in the present state. Consider firstly the general case of the process with denumerable set of states $E = \{1, 2, \dots\}$. To describe the object behavior by a Markov process let us introduce an enlarged states space $\mathcal{E} = E \times [0, \infty)$ and consider two dimensional process $Z(t) = \{S(t),\, X(t)\}$, where the first component $S(t) \in E$ shows the object's state, and the second one $X(t) \in [0, \infty)$ denotes the time spent in the state since the last entrance into it. Denote by $\alpha_i(x)$ and $\beta_i(x)$ $(i \in E)$ the transition intensities from the state i to the states $i+1$ and $i-1$ respectively under the condition that the time spent at the state i equals to x.

Remark. If the stay time at the state i is considered as a minimum of two independent random variables (r.v): time A_i till to transition into the "next" state $i+1$ and time B_i till to transition into the "previous" state $i-1$ with cumulative distribution functions (c.d.f.) $A_i(x)$, $B_i(x)$, then the introduced process can be considered as a special case of semi-Markov process (SMP) (see Korolyuk and Turbin (1976)), with conditional transition p.d.f.'s $\alpha_i(x)$ $\beta_i(x)$. Nevertheless, the given formalization open the new possibilities for the investigations and moreover in the degradation models we are studying the conditional probability state distribution on given life period, that was not

investigated previously.

Denote by $\pi_i(t,x)$ the p.d.f. of the process $Z(t)$ at time t,

$$\pi_i(t,x)dx = \mathbf{P}\{S(t)=i,\ x \le X(t) < x+dx\}.$$

These functions satisfy to the Kolmogorov's system of differential equations

$$\frac{\partial \pi_i(t,x)}{\partial t} + \frac{\partial \pi_i(t,x)}{\partial x} = -(\alpha_i(x)+\beta_i(x))\pi_i(t,x), \qquad 0 \le x \le t < \infty,\ \ i \in E \tag{75.2.1}$$

with the initial and boundary conditions

$$\begin{cases} \pi_1(t,0) &= \delta(t) + \int\limits_0^t \pi_2(t,x)\beta_2(x)dx, \\ \pi_i(t,0) &= \int\limits_0^t \pi_{i-1}(t,x)\alpha_{i-1}(x)dx + \int\limits_0^t \pi_{i+1}(t,x)\beta_{i+1}(x)dx, \quad i \in E. \end{cases} \tag{75.2.2}$$

In the following we will suppose the process to be non reducible, non degenerated. The conditions for this in terms of SMP might be found for example in McDonald (1978) and Jacod (1971). For the non reducible, non degenerated generalized B & D process the Kolmogorov's system of equations (75.2.1) with initial and boundary conditions (75.2.2) has a unique solution over all time axis.

It is possible to show by the method of characteristics Petrovsky (1952) that its solution has a form

$$\pi_i(t,x) = g_i(t-x)(1-A_i(x))(1-B_i(x)), \qquad 0 \le x \le t < \infty,\ \ i \in E, \tag{75.2.3}$$

where the functions $g_i(t)$ in accordance with the initial and boundary conditions (75.2.2) satisfy to the system of equations

$$\begin{cases} g_1(t) &= \delta(t) + \int\limits_0^t g_2(t-x)(1-A_2(x)b_2(x))dx, \\ g_i(t) &= \int\limits_0^t g_{i-1}(t-x)a_{i-1}(x))(1-B_{i-1}(x))dx + \\ &+ \int\limits_0^t g_{i+1}(t-x)(1-A_{i+1}(x)b_{i+1}(x))dx, \qquad i = 2,3,\ldots. \end{cases} \tag{75.2.4}$$

The form of these equations shows that their solution should be found in terms of its Laplace transforms (LT's). Therefore by passing to the LT's with respect to both variables into relations (75.2.3) after the change of the integration order one can get

$$\tilde{\tilde{\pi}}_i(s,v) \equiv \int\limits_0^\infty e^{-st} \int\limits_0^t e^{-vx}\pi(t,x)\,dx\,dt = \tilde{g}_i(s)\tilde{\gamma}_i(s+v), \tag{75.2.5}$$

where $\tilde{g}_i(s)$ are the LT of the functions $g_i(t)$, and the functions $\tilde{\gamma}_i(s)$ are

$$\tilde{\gamma}_i(s) = \int_0^\infty e^{-st}(1 - A_i(t))(1 - B_i(t))dt.$$

From the other side by passing to the LT's with respect to variable t in the system (75.2.4) one get

$$\left\{ \begin{array}{rcll} \tilde{g}_1(s) - \tilde{g}_2(s)\tilde{\psi}_2(s) & = & 1, & \\ -\tilde{g}_{i-1}(s)\tilde{\phi}_{i-1}(s) + \tilde{g}_i(s) - \tilde{g}_{i+1}(s)\tilde{\psi}_{i+1}(s) & = & 0, & i = 2, 3 \dots \end{array} \right. \tag{75.2.6}$$

where the functions $\tilde{\phi}_i(s)$ and $\tilde{\psi}_i(s)$ are given by the relations

$$\tilde{\phi}_i(s) = \int_0^\infty e^{-sx} a_i(x)(1-B_i(x))dx, \quad \tilde{\psi}_i(s) = \int_0^\infty (1-A_i(x))b_i(x))dx, \quad i = 1, 2, \dots.$$

In the case of finite number $n+1$ of states in the above system one should put $\tilde{\phi}_{n+1}(s) = 0$.

The closed form solution of this system in general case even in the simplest case of usual B & D process does not possible. In spite of the above equations not being possible to solve in closed form they provide calculation different characteristics of the process. Consider some of them.

75.2.2 Stationary probability distribution

For calculation of the process $Z(t)$ macro-states stationary probabilities

$$\pi_i = \lim_{t\to\infty} \pi_i(t) = \lim_{t\to\infty} \int_0^t \pi(t, x)dx = \lim_{t\to\infty} \int_0^t g_i(t - x)(1 - A_i(x))(1 - B_i(x))dx$$

we use the connection between asymptotic behavior of functions at infinity and their LT's at zero. Letting $\tilde{\gamma}(0) = \gamma$ and taking into account that accordingly to (75.2.5) $\tilde{\pi}_i(s) = \tilde{\tilde{\pi}}_i(s, 0)$, we find

$$\lim_{t\to\infty} \pi_i(t) = \lim_{s\to 0} s\tilde{\pi}_i(s) = \gamma_i \lim_{s\to 0} s\tilde{g}_i(s).$$

Thus, for the problem solution it is necessary to calculate the values

$$g_i = \lim_{s\to 0} s\tilde{g}_i(s),$$

for what we use equations (75.2.6). By multiplying these equations by s, and by passing to limit when $s \to 0$ in them and taking into account that $\phi_i + \psi_i = 1$ we get the recursive relations

$$g_i\phi_i - g_{i+1}\psi_{i+1} = g_{i-1}\phi_{i-1} - g_i\psi_i, \qquad i = 2, 3 \dots, \tag{75.2.7}$$

where the notations $\phi_i = \tilde{\phi}_i(0)$, $\psi_i = \tilde{\psi}_i(0)$ were used. Now because $\phi_1 = 1$ from it follows from the first of equations (75.2.6) that $g_1\phi_1 - g_2\psi_2 = 0$. With the help of the last recursive relation it is possible to calculate coefficients g_i and find the stationary distribution, that is given in the following theorem.

Theorem 1. *For the generalized B & D process stationary regime existence it is necessary the convergence of the series*

$$g_1^{-1} = \sum_{1 \le i < \infty} \gamma_i \prod_{1 \le j \le i} \frac{\phi_{j-1}}{\psi_j} < \infty. \tag{75.2.8}$$

In this case the stationary probabilities are given by the formula

$$\pi_1 = g_1\gamma_1, \qquad \pi_i = g_1\gamma_i \prod_{1 \le j \le i} \frac{\phi_{j-1}}{\psi_j}, \quad i = 2, 3, \ldots. \qquad \heartsuit \tag{75.2.9}$$

Moreover, from the form of stationary probabilities it follows the next important corollary

Corollary. *The macro-states stationary probabilities of generalized B & D process are insensitive to the shape of distributions $A_i(x)$, $B_i(x)$ and depend on r.v. A_i, B_i and their distributions only by means of probabilities of jumps embedded random walk up and down and mean time of the process stay in the given state,*

$$\phi_i = \mathbf{P}\{A_i \le B_i\}, \qquad \psi_i = \mathbf{P}\{A_i > B_i\}, \qquad \textit{and} \quad \gamma_i = \mathbf{E}[\min A_i, B_i]. \qquad \heartsuit \tag{75.2.10}$$

For the process with finite number of states $n+1$ in the Kolmogorov's system of equations (75.2.1) one should put $\alpha_{n+1}(x) \equiv 0$. In this case the stationary probabilities have the same form (75.2.9), but the normalizing constant (75.2.8) should be changed by an appropriate finite sum. In the case of exponential distributions $A_i(x) = 1 - e^{-\alpha_i t}$ and $B_i(x) = 1 - e^{-\beta_i t}$ the formulas (75.2.9) are reduced to the stationary probabilities of the usual B & D process.

75.2.3 Distribution of the process states given on life period

For many phenomenons especially for degradation processes more appropriate is absorbing process model. For the generalized B & D process with absorbing state $n + 1$ in the Kolmogorov's system of equations (75.2.1) one should put $\alpha_{n+1}(x) = \beta_{n+1}(x) \equiv 0$. In this case the equation for the function $\pi_{n+1}(t)$ takes the form

$$\frac{\partial \pi_{n+1}(t,x)}{\partial t} + \frac{\partial \pi_{n+1}(t,x)}{\partial x} = 0, \tag{75.2.11}$$

with the initial and boundary condition

$$\pi_{n+1}(t,0) = \int_0^t \pi_n(t,x)\alpha_n(x)dx. \qquad (75.2.12)$$

Thus, all functions $\pi_i(t,x)$ $(i = \overline{1,n})$ have the same solution (75.2.3) as before. But the function $\pi_{n+1}(t,x)$, being a constant over the characteristics, is $\pi_{n+1}(t,x) = g_{n+1}(t-x)$, where due to the boundary conditions (75.2.12) it follows that

$$g_{n+1}(t) = \int_0^t g_n(t-x)a_n(x)(1-B_n(x))dx.$$

The solution $\pi_i(t)$ of the system of equations (75.2.1, 75.2.2, 75.2.11, 75.2.12) gives the probability of the object to be in some state jointly with its life period T,

$$\pi_i(t) = \mathbf{P}\{S(t) = i,\ t < T\}, \qquad i = 1,2,\ldots,n.$$

For the degradation problems investigation more useful and adequate characteristic is the conditional state probability distribution on given object's life period for which the following representation is true

$$\bar{\pi}_i(t) = \mathbf{P}\{S(t) = i \mid t < T\} = \frac{\pi_i(t)}{R(t)}, \qquad i = 1,2,\ldots,n$$

where $R(t) = 1 - \pi_{n+1}(t)$ is the reliability (survival) function of the object. For its LT $\tilde{R}(s)$ one can find

$$\tilde{R}(s) = \frac{1}{s} - \frac{1}{s}\tilde{g}_n(s)\tilde{\phi}_n(s) = \frac{1}{s}(1 - \tilde{g}_n(s)\tilde{\phi}_n(s)). \qquad (75.2.13)$$

We are interesting in limits of the conditional probabilities states given life period. To calculate their we should evaluate the asymptotic behavior of the functions $\pi_i(t)$ $(i = 1,2,\ldots,n)$ and $R(t)$ when $t \to \infty$. We will do that with the help of their LT. Denote by $\Delta(s)$ the determinant of the matrix of coefficients of n first equations of the system (75.2.6) and by $\Delta_i(s)$ the determinant of the same matrix in which i-th column is changed by the vector-column of the equation right side (vector e_n). Then taking into account the expression (75.2.5), the LT of the function $\pi_{n+1}(t)$, and the solution of the system (75.2.6) in terms of the Kramer's rule, we get

$$\begin{cases} \tilde{\pi}_i(s) &= \tilde{\gamma}_i(s)\tilde{g}_i(s) = \tilde{\gamma}_i(s)\frac{\Delta_i(s)}{\Delta(s)} \qquad (i = \overline{1,n}), \\ \tilde{\pi}_{n+1}(s) &= \frac{\tilde{\phi}_n(s)}{s}\tilde{g}_n(s) = \frac{\tilde{\phi}_n(s)\Delta_n(s)}{s\Delta(s)}. \end{cases}$$

Theorem 2. *Asymptotical behavior of the functions* $\pi_i(t)$ *and* $R(t)$ *when* $t \to \infty$ *coincide and is determined by the maximal non-zero root* s_1 *of the characteristic equation* $\Delta(s) = 0$. *This provide the existence of the limit*

$$\bar{\pi}_i = \lim_{t\to\infty} \bar{\pi}_i(t) = \frac{\tilde{\pi}_i(s_1)}{\tilde{R}(s_1)}.$$

Proof. The proof is based on the fact, that the asymptotic behavior of the functions $\pi_i(t)$ and $R(t)$ when $t \to \infty$ is coincide, that follows from their LT analysis. ♡

To illustrate the above results a system with only three states, which can be considered as an example of the aggregated states model (see Korolyuk and Turbin (1978), Korolyuk, and Korolyuk (1999)), where all states of each group: normal functioning N, degradation D, and failure F are joined into one was considered.

75.3 Conclusion

Generalized Birth & Death Processes, which are special class of Semi-Markov Processes are introduced for modelling the degradation processes. The special parametrization of the processes allows to give more convenient presentation of the results. The special attention is focused to the conditional state probabilities given life cycle, which are the mostly interesting for the degradation processes.

References

1. A. Lisniansky, A. and Levitin G. (2003). *Multi-State System Reliability. Assessment, Optimization and Application*, World Scientific, New Jersey, London, Singapore, Hong Kong, 358p.
2. Dimitrov, B. Rykov, V. and Stanchev P.(2002). On Multi-State Reliability Systems. In: *Proceedings MMR-2002.* Trondheim (Norway) June 17-21, 2002.
3. Rykov, V., Dimitrov B. (2002). On Multi-State Reliability Systems. In: *Applied Stochastic Models and Information Processes.* Proceedings of the International Seminar, Petrozavodsk, Sept. 8-13. Petrozavodsk, 2002, pp. 128-135. See also http://www.jip.ru/2002-2-2-2002.htm
4. Rykov, V., Dimitrov, B., Green Jr., D., Snanchev . (2004) Reliability of complex hierarchical systems with fault tolerance units. In: *Proceedings MMR-2004.* Santa Fe (U.S.A.) June, 2004. (Printed in CD).

5. Dimitrov, B., Green Jr., D., Rykov, V, and Stanchev P. (2004). Reliability Model for Biological Objects. In: *Longevity, Aging and Degradation Models. Transactions of the First Russian-French Conference (LAD-2004)*, Saint Petersburg , June 7-9, 2004, Ed. by V. Antonov, C. Huber, M. Nikulin, V. Polischook, Saint Petersburg State Politechnical University, SPB, 2004. Vol. 2, pp. 230-240.
6. Rykov, V., Efrosinin D. (2004) Reliability Control of Fault Tolerance Units. In: *Abstracts of The 4-th International Conference on Mathematical Methods in Reliability (MMR-2004)*, Santa Fe, (USA), 21-25 July, 2004 (Published in CD).
7. Rykov, V, Efrosinin D. (2004). Reliability Control of of Biological Systems with failures. In: *Longevity, Aging and Degradation Models. Transactions of the First Russian-French Conference (LAD-2004)*, Saint Petersburg , June 7-9, 2004, Ed. V.Antonov, C.Huber, M.Nikulin, V.Polischook, Saint Petersburg State Politechnical University, SPB, 2004. Vol. 2, pp. 241-255.
8. Rykov, V., Buldaeva E. (2004). On reliability controlof fault tolerance units: regenerative approach. In: *Transactions of XXIV International Siminar on Stability Problems for Stochastic Modes*, September 10-17, 2004, Jurmala, Latvia. Transport and Telecommunication Institute, Riga, Latvia, 2004.
9. Korolyuk, V.S., Turbin A.F. (1976) *Semi-Markov processes and their applications.* Kiev: "Naukova dumka". 1976, 184p. (in Russian)
10. McDonald, D. (1978). On semi-Markov and semi-regenerative processes. I, II. *Z. fur Wahrch. verw.Geb.*, 42 (1978), No. 2, pp. 261-377; *Ann. of Prob.*, 6 (1978), No.6, pp. 995-1014.
11. Jacod, J. (1971) Theorems de renouvellement et classification pour les chaines semi-Markoviennes. *Ann. inst. Henri Poincare*, sect. B, 7 (1971), No.2, pp.83-129.
12. Korolyuk, V.S., Turbin A.F. (1978). *Phase aggregation of complex systems.* Kiev: "Vish shkola". 108p. (in Russian).
13. Korolyuk, V.S., Korolyuk V.V. (1999). *Stochastic Models of Systems.* Kluver Academic Publishers.
14. Petrovsky, I.G. (1952). *Lectures on the theory of usual differential equations* . M.-L.: GITTL. 1952, 232p. (in Russian).
15. Feller, W. (1966). *An Introduction to Probability Theory and its Applications V. II* John Wiley & Sons, Inc. N.Y.-Lnd.-Sidney, 1966.
16. Lavrent'ev, M.A., Shabat B.V. (1958). *Methods of the theory of complex variable functions.* M.: Phismatgiz. 1958, 678p.

76

The System of Cerebral Circulation: An Assessment of its State with Cross-Spectral Analysis

Semenyutin V.B., Aliev V.A, Patzak A.†, Kozlov A.V., Nikitin P.I.

Russian Polenov Neurosurgical Institute, St. Petersburg, Russia
†Johannes-Mueller Institute of Physiology University Hospital Charité, Humboldt-University of Berlin, Germany

76.1 Introduction

It has been established, that spontaneous changes of systemic blood pressure (BP) and blood flow velocity (BFV) are characterized by four main wave processes: heart rate (0.5-1.6 Hz), respiratory excursions (0.15-0.65 Hz), Meyer's waves (0.05-0.15 Hz) and B-waves (0.008-0.05 Hz)[2, 3, 6].

Impaired cerebral autoregulation (CA) is one of the leading links in pathogenesis of disorders occurring in the system of cerebral circulation (SCC). CA is a multi-component mechanism, ensuring stability of cerebral blood flow not only in step changes of BP, but also in its spontaneous oscillations within the middle-frequency range. From this point of view CA is considered to be a filter system, normally transmitting only high-frequency oscillations (0.15-0.5 Hz) of BP, which are characterized by high coherence and a smaller phase shift (PS) as compared with analogous oscillations of BFV. At the same time this system damps Meyer's waves (M-waves) of BP. It results in low coherence and a large PS between BP and BFV. The obtained data [3-6] reflect dependence of CA on frequency. Thus, it is more effective within the middle-frequency range rather than the high-frequency one. Disorders of CA lead to an increase of a transmitting capacity of the filter and, as a result, to higher coherence and PS between BP and BFV within the range of M-waves.

Information on B-waves is contradictory. B-waves are most probable to reflect a state of mechanisms, responsible for regulation of SCC and mediated by smooth-muscle cells of cerebral vessels or stem pacemakers, which change cerebral blood flow with a certain rhythm by effecting activity of vasomotor neurons.

It should be noted, that world literature sticks up for the conception of

possible use of spectral analysis for estimation of CA. However, there are many contradictions from a methodological point of view. Different authors use spectral analysis for search after new informative parameters, allowing to estimate a state of CA. It causes development of special programs of mathematical analysis (including spectral analysis), aimed at solution of definite problems. Today some parameters of cross-spectral analysis have already been determined (PS, coherence, gain). They can be calculated with the help of any of the above programs [4, 7-9] and used for routine diagnosis of normal and pathologic states of CA. However, spectral analysis is used mainly in clinics with specialized research centers, studying CA on the fundamental basis. From our point of view, one of the causes, hampering wide introduction of this method into clinical practice, is absence of available program which would allow to calculate PS, coherence of slow oscillations of BP and BFV.

One of the most spread programs, used for analysis of data, is "Statistica for Windows". It permits to solve general statistical problems (parametric and non-parametric statistics, regression, discriminant analysis), which are of great importance for conclusive medicine. Besides, it makes it possible to analyze time series with the help of an autoregression model, an autoregressive moving average model, fast Fourier transformation in compliance with a standard algorithm of operating mass data.

The goal of the present research consisted of estimation of cerebral hemodynamics with cross-spectral analysis and use of "Statistica 6.0 for Windows".

76.2 Materials and Methods

The study was carried out on 30 healthy volunteers and 50 patients with different cerebrovascular pathology. The age of volunteers varied from 18 up to 51 years. They had normal blood pressure and HR; there were no chronic and acute cardiac or cerebral pathology in their life history. The age of patients varied from 18 up to 64 years. Intracranial aneurysms, arteriovenous malformations (AVM) and carotid-cavernous anastomosis were watched in 39, 10 and 1 cases respectively. The study was carried out both in an acute stage of the disease and during a long-term period.

BFV in the middle cerebral artery was monitored with the Multi Dop X system (DWL, Germany). BP was recorded, using a non-invasive method with the Finapres-2300 apparatus (Ohmeda, USA). During monitoring a person was in a supine position with his head lifted up to 30^{o}. Continuous recording was carried out during 5 min. It was done at rest and spontaneous breathing.

CA was estimated by calculating the rate of regulation (RoR) with a cuff test [1]. This method lies in analysis of changes of cerebrovascular resistance in response to acute reduction of BP, achieved by transient hyperemia in the lower extremities, which follows thigh decompression with pneumatic cuffs.

Spectral analysis was carried out, using standard procedures by "Statistica 6.0 for Windows" program. PS between BFV and BP within the M-waves range, B-wave spectral density (BWSD) and B-wave amplitude (BWA) were estimated. While carrying out spectral analysis we proceeded from potentialities of digital processing of data by the Multi Dop X system. A time series at an interval of 282 s was chosen for studying the spectrum of BFV. According to the theorem of Kotelnikov-Shannon, estimation of the spectrum of low-frequency oscillations demands analysis of a time series for a period, which is two times longer than a maximum period of low-frequency oscillations (it is 120 s for B-waves). A time series with duration of 282 s is optimum, as M- and B-waves are represented by a sufficient number of harmonics, which allows to carry out detailed simultaneous spectral analysis in both ranges.

Statistical processing of data was based on application of standard methods. Parametric (Student t-criterion) and non-parametric (Kholmogorov-Smirnov criterion, Wilcoxon criterion, Mann-Withney criterion) criteria were used. The difference was considered to be reliable in $p < 0.05$.

76.3 Results

Mean values of spontaneous oscillations of systemic and cerebral hemodynamics, RoR and cross-spectral analysis are given in Table 1.

BWSD of BFV did not exceed 1000 $(\text{cm/s})^2/\text{Hz}$ in healthy individuals with normal values of BFV and BP. PS between BFV and BP in the M-wave range correlated with normal values of RoR and was within the limits of 1 rad (57^{o}). It confirmed a normal state of CA and agreed with data of authors, who used another statistical programs of data processing [3, 5, 9].

As for patients with intracranial aneurysms in a remote period of hemorrhage, BWSD and BWA were practically identical to analogous values, watched in healthy individuals. Probably, it could be explained by relative stabilization of a whole SCC. At the same time values of PS and RoR were reliably lower, than normal indices. It was indicative of preserved disorders of CA in this group of patients, but there was a tendency to their normalization. In a hemorrhagic period (i.e. at the stage of lesions of SCC, marked to the utmost) patients were characterized by considerable increase of BWSD and BWA and more severe disorders of CA, manifesting themselves in reliably low values of PS and RoR.

Cases with moderate vasospasm or its absence, observed in a hemorrhagic period, had mean values of BWSD and BWA, which were reliably lower in comparison with patients with severe or critical vasospasm; values of PS and RoR were reliably higher (Table 1).

Severe and critical vasospasm is one of the most serious complications of aneurysm rupture, determining a course of an early postoperative period after open intracranial interventions to a considerable extent. Data of preoperative

Table 1. Mean values of BP, HR, BFV, RoR and crosspectral analysis for a period of 282 s in volunteers and different groups of patients

Groups of volunteers and patients	Data of systemic and cerebral homodynamics						
	Mean values				Data of crosscpectral analysis		
	BP [mmHg]	HR [min^{-1}]	BFV [m/s]	RoR [%/s]	PS [rad]	BWSD [(m/s)^2/Hz]	BWA [m/s]
Volunteers (n=30)	90±3	82±1	68±3	34±5	1.0±0.1	557±70	3.0±0.2
Cerebral aneurysms (n=39)							
>30 days after SAH (n=26)	89±3	83±1	62±3	20±1	0.6±0.1	520±33	3.5±0.2
1-21 days after SAH (n=13)	94±3	78±2	120±17	13±1	0.4±0.1	2363±830	5.8±0.9
0-1 grades of VS (n=6)	90±5	84±2	65±7	14±1	0.5±0.1	605±247	3.8±0.7
2-3 grades of VS (n=7)	99±4	69±3	198±11	9±3	0.2±0.1	4875±1595	8.8±1.2
AVM (n=10)							
Pathological side	89±3	84±2.3	131±9	8±1	0.5±0.1	429±103	2.5±0.3
Contralateral side			67±6	21±1	0.9±0.1	413±118	1.9±0.4

monitoring of BFV, analyzed retrospectively, demonstrated reliable difference of indices of its BWSD in cases with severe and critical vasospasm and its dependence on a course of an early postoperative period (Table 1).

BWSD of patients with early development of postoperative neurological complications (7091 ± 2235 $(\mathrm{cm/s})^2/\mathrm{Hz}$) was reliably higher ($p < 0.05$) than that of cases without complications (1921 ± 439 $(\mathrm{cm/s})^2/\mathrm{Hz}$).

Estimation of a PS between BFV and BP within the M-wave range, carried out in patients with AVM in a preoperative period, revealed its reliable asymmetry on the side of AVM localization and the contralateral intact side. Low values of PS on the AVM side appear to be conditioned by a degree, to which a shunting process is marked, and participation of an afferent vessel in feeding intact area of the brain, adjacent to AVM (Table 1). Thus, one can make the following supposition: the higher RoR and a value of PS, the greater contribution of an afferent vessel to blood supply of cerebral areas, adjacent to AVM, and its functional significance. It results in a higher risk of neurological complication in radical embolization of AVM.

Changes of BWA were watched at different stages of endoavascular interventions, performed in patients in a remote period of hemorrhage and craniocerebral trauma. Depending on a course of an intraoperative period, all the cases

were divided into two groups: patients without complications and patients with developed intraoperative neurological complications.

Mean values of BP, BFV and BWA, watched in cases at various stages of endovascular intervention, are given in Table 2.

Changes of BFV and BP, irrespective of presence of complications in a perioperative period, were insignificant. The patients of both groups had spontaneous breathing. A preoperative value of BWA was considerably higher ($p < 0.05$) in cases with complications. At the same time there were no objective signs of symptoms augmentation. The main stage of operation was characterized by further increase of BWA on both sides, which was accompanied by development or augmentation of neurological symptoms. There was inconsiderable reduction of BWA after completion of operation (coil and balloon detachment, glue administration, catheter removal). There was no regression of neurological symptoms.

Table 2. Mean values of BP, BFV and BWA on the different stages of endovascular intervention

Groups of patients and stages of operation	BP [mm Hg]	BF parameters	
		BFV [cm/s]	BWA of BFV [cm/s]
Patients without complications (n=6)			
Preoperative	79±9	77±11	3.9±0.6
Intraoperative	84±10	85±12	7.7±1.1
Postoperative	84±11	87±14	4.2±0.8
Patients with complication (n=6)			
Preoperative	85±11	71±13	9.6±1.1
Intraoperative	96± 9	89±18	12.1±2.6
Postoperative	90± 7	66±12	10.4±2.9

76.4 Discussion

The cross-spectral analysis of BFV and BP carried out with the help of "Statistica for Windows" allowed getting mean values of BWSD and BWA, as well as PS between BFV and BP within the M-wave range in healthy individuals. They were analogous to the results, obtained by different authors, who used other statistical programs [3, 5, 6]. Indices of PS correlated with parameters of CA, estimated with a cuff test (RoR). BWSD values of less than 1000 $(\text{cm/s})^2/\text{Hz}$, ascertained with the help of classification trees, were indicative of a normal state of stem structures, ensuring adequate functioning of SCC. The above advantages of spectral analysis make it a preferable method in studying SCC not only in healthy individuals, but also in cases with different neurosurgical pathology.

Patients with intracranial aneurysms, subject to open operations, had inter- and intra-group difference in BWSD, BWA and PS within the M-wave range. Dependence of BWSD on a degree of vasospasm is of peculiar importance. It

is known, that determining indications for intracranial operations in a hemorrhagic period one should consider severity of a patient's state, as well as presence and a degree of vasospasm, diagnosed on the basis of angiographic and Doppler examination. Sometimes data of Doppler examination play a decisive role, as operations, performed in patients with severe or critical vasospasm, are accompanied by development of early postoperative neurological complications more frequently in comparison with interventions in cases with moderate vasospasm or its absence. Indications for an operation can be determined more precisely, using results of BWSD estimation, which turned out to be much higher in cases with postoperative neurological complications than in cases without postoperative neurological complications. It should be noted, that there was no reliable difference in BFV, PS between BFV and BP, RoR on the side of vasospasm localization in two groups.

This fact allows to suppose, that there can be subsequent disorders of static (according to values of a PS within the M-wave range) and dynamic (according to results of a cuff test) CA, watched in SCC in an acute hemorrhagic period. They can stabilize or regress against a background of treatment or involve stem structures, regulating cerebral blood flow, into a pathologic process (increase of BWSD and BWA) and form a vicious circle. Thus, values of BWSD and BWA can be used as an additional criterion in choosing tactics of treatment, i.e. an urgent operation with the purpose of preventing repeated hemorrhage or conservative intensive care, aimed at reduction of BWSD and BWA, for improvement of a patient's initial preoperative state.

Considerable increase of BWA was observed in development of intra- and postoperative neurological complications in cases, subject to intravascular interventions in a remote hemorrhagic period. In such situations a growth of BWSD and BWA may be conditioned by reflex spasm in short perforating branches, supplying the brain stem. This spasm is a response to manipulations, performed during an approach to a pathologic substrate through major extra- and intracranial vessels, and direct contact with it. Timely detection of increased BWSD and BWA would allow to prognosticate intra- and postoperative neurological complications in intravascular interventions and to take adequate measures for their prevention.

As for patients with AVM, information on a state of CA, obtained on the basis of estimation of PS between BFV in an afferent vessel and BP in the M-wave range, can be used for determination of its functional significance. It will help to assess possibility of radical exclusion of AVM from circulation more precisely. Today evaluation of functional significance is based on an invasive method and consists in measuring blood pressure in an afferent vessel, volumetric blood flow velocity in an afferent vessel and using pharmacologic tests. According to our opinion, functional insignificance of afferent vessels is characterized by reduction of RoR and, as a result, by a small PS, which can be

determined before operation. Thus, probability of safe exclusion of AVM can be prognosticated with a simpler and non-invasive method. However, confirmation of this supposition demands comparison of results of estimating PS and blood pressure in an afferent vessel, which is a subject of an independent study.

76.5 Conclusion

We tried to reflect modern views on SCC, formed on the basis of a rather new, but prospective method of estimation of cerebral hemodynamics, introduced into clinical practice, i.e. spectral analysis of spontaneous oscillations of BP and BFV. Thus, this carrying out spectral analysis of spontaneous oscillations of BFV and BP, which includes successive determination of BWSD and BWA, PS between BFV and BP within the M-wave range can be used for non-invasive estimation of SCC.

References

1. Aaslid R., Lindegaard K.F., Sorteberg W., Nornes H. Cerebral autoregulation dynamics in humans, *Stroke*, 1989, **20, N.1**, 45-52.
2. Birch AA, Dirnhuber MJ, Hartley-Davies R, Iannotti F, Neil-Dwyer G. Assessment of autoregulation by means of periodic changes in blood pressure, *Stroke*, 1995,**26**, 834-837.
3. Diehl RR, Linden D, Lucke D, Berlit P. Phase relationship between cerebral blood flow velocity and blood pressure. A clinical test of autoregulation *Stroke*, 1995, **26**, 1801-1804.
4. Giller C.A. The frequency-dependent behavior of cerebral autoregulation, *Neurosurgery*, 1990, **27**, 362-368.
5. Haubrich C. Wendt A., Diehl R.R., Kl?tzsch C. Dynamic Autoregulation Testing in the Posterior Cerebral Artery *Stroke*, 2004, **35**, 848.
6. Lang E.W., Diehl R.R., Mehdorn H.M. Cerebral autoregulation testing after aneurysmal subarachnoid hemorrhage: the phase relationship between arterial blood pressure and cerebral blood flow velocity, *Crit Care Med*, 2001, **29**, 158-163.
7. Panerai R.B., Rennie J.M., Kelsall A.W., Evans D.H. Frequency-domain analysis of cerebral autoregulation from spontaneous fluctuations in arterial blood pressure *Med Biol Eng Comput*, 1998, **36**, 315-322.
8. Reinhard M., Roth M., Muller T., Czosnyka M., Timmer J., Hetzel A. Cerebral autoregulation in carotid artery occlusive disease assessed from spontaneous blood pressure fluctuations by the correlation coefficient index *Stroke*, 2003, **35**, 2138-2144.
9. Zhang R., Zuckerman J.H., Giller C.A., Levine B.D. Transfer function analysis of dynamic cerebral autoregulation in humans *Am J Physiol.*, 1998, **274**, H233-H241.

77

Dynamic Modeling of Greek Life Table Data

Christos H. Skiadas

Technical University of Crete, Data Analysis and Forecasting Laboratory

Abstract: In this paper we formulate and apply a dynamic model expressing the human life table data by using the first-passage-time theory for a stochastic process. The model is based on the First Exit Time theory and is applied to the mortality data in Greece. A stochastic simulation is also performed for the Health State Function proposed and the related stochastic paths. Furthermore the implications of the proposed model and the results derived are discussed.

Keywords and phrases: Life table data, Stochastic model, First passage time densities.

77.1 Introduction

In previous studies (see Janssen and Skiadas (1995)) the concept of health state was modeled by a continuous-time stochastic process. This means that they started from a stochastic process

$$S = (S(t), t \geq 0)$$

defined on a completed filtered probability space

$$(\Omega, \theta, (\theta_t), P)$$

where the random variable $S(t)$ represents the health state of an individual at time t. The event 'death' is defined as the time that this state health hits for the first time a minimal health level called a. Consequently, the life duration of the individual is the value of this hitting time T of the set $(0, a)$ for the process S.

In this paper, we introduce briefly the general theory of dynamic models for modelling the human life and we present a particularly good model fitting quite well the Greek mortality tables.

77.2 The Model

77.2.1 The stochastic model and the related parameters

The proposed model is a stochastic model of continuous time provided by the stochastic differential equation:

$$dS(t) = \mu(t)dt + \sigma dW(t)$$

where $S(t)$ is a stochastic variable expressing the state of health of an individual, $\mu(t)$ a function of time expressing the infinitesimal mean development of the state of human health, σ the infinitesimal variance of the human health that is assumed to be constant in the proposed model and $W(t)$ the standard Wiener process. $S(t)$ is easily obtained by direct integration from the above stochastic differential equation. Then $S(t)$ is given by:

$$S(t) = S_0 + \int_{t_0}^{t} \mu(s)ds + \sigma[W(t) - W_0]$$

where S_0 is the value of $S(t)$ at time $t = 0$.

Now our main task is to obtain an analytic form of the function $\mu(t)$. Consider that the mean value of $S(t)$ is a function $H = H(t)$ given by:

$$H(t) = E[S(t)] = S_0 + \int_{t_0}^{t} \mu(s)ds$$

The function $\mu(t)$ is related to $H(t)$ by the formula:

$$\mu(t) = \frac{\vartheta H(t)}{\vartheta t}$$

We expect that the mean value H or the mean state of health function during lifetime will follow a general growth and decline path. There is a fast improvement of the state of human health after birth when $H(t)$ is close to a low level. Then it follows a period of slow improvement and then decline. The function $\mu(t)$ as the derivative of $H(t)$ must begin from very high values in the beginning for time t close to zero and decline continuously reaching zero when $H(t)$ is at the maximum and then takes negative values. This function, as it must have negative derivative ($d\mu/dt < 0$), expresses the inevitable decline of the infinitesimal mean of the state of human health and is proposed to be called as the *Mortality Function.*

The main assumption regarding the function $\mu(t)$ is that this function may express two distinct time-processes related to the state of human health. The one is the first period of life-time modelled by a function $u = u(1/\sqrt{t})$ that must be a fast decreasing function of time expressing the fast improvement of the

state of human health during the first years after birth. The other is a function $v = v(t)$ that expresses the gradually increasing and then decreasing state of health during the total period of the life-time. The function $\mu(t)$ is then given by:

$$\mu(t) = u(\frac{1}{\sqrt{t}}) + v(t)$$

An approximation of $\mu(t)$ is achieved by expanding $u(t)$ and $v(t)$ in Taylor-Maclaurin series and retaining the first two terms to the right for the fast decreasing function u and four terms to the right for the function v. The two series' expansions give the form:

$$v(t) = a_0 + 2a_1t + 3a_2t^2 + 4a_3t^3$$

$$u(\frac{1}{\sqrt{t}}) = b_0 + \frac{b_1}{2\sqrt{t}}$$

For both cases it is assumed that the parameters a_0 and b_0 are equal to zero, because close to zero they don't contribute much to the function $\mu(t)$ that takes very high values due to the second term to the right of u. Then the selected approximation for $\mu(t)$ has the following form:

$$\mu(t) = 2a_1t + 3a_2t^2 + 4a_3t^3 + \frac{b_1}{2\sqrt{t}}$$

Now $H(t)$, the mean value of $S(t)$, under the assumptions $t_0 = 1$ and $S_0 = 0$ is as follows:

$$H(t) = E[S(t)] = \int_1^t \mu(s)ds = a_1t^2 + a_2t^3 + a_3t^4 + b_1\sqrt{t} - c$$

where

$$c = a_1 + a_2 + a_3 + b_1$$

The selection of the starting time $t_0 = 1$ is done because our data series in life tables for the deaths at age between x and $x+1$ denoted by $d(x)$ begin from the first year when t equals 1.

77.2.2 The hitting time density function

The main assumption regarding the state of human health is that it follows a stochastic process expressed by $S(t)$ and that the end of the life-time is reached when the stochastic variable $S(t)$ arrives at a minimum level of health state here denoted by a. This level a in terms of the first passage time theory is expressed by a single barrier located at a distance a from the origin. Then, the density function $g(t)$ expressing the distribution of the first passage from the barrier a

is exactly the probability density function that provides the number of deaths between t and $t + dt$ where t is nothing else but the age of the individuals.
After the pioneering work of Siegert (1951) regarding analytic solutions of the first-passage-time probability problem a quite extensive bibliography has appeared in the last decades. The aim was either to derive analytic solutions or, when this is not possible, to approximate satisfactorily the first-passage-time density function under consideration. In most cases the task achieved by using an integral equation of the Volterra type (see Buonocore *et al.* (1987) and Giorno *et al.* (1990)) of the form:

$$g(\lambda(t), t; x_0) = -2\psi(\lambda(t), t; x_0) + 2\int_0^t d\tau \psi(\lambda(t), t; \lambda(\tau), \tau) g(\lambda(\tau), \tau; x_0)$$

where $x_0 < \lambda(0) \equiv a$ where $\lambda(t)$ denotes a smooth function of time that cuts the stochastic paths provided by the density function $g(t)$.
Another approach (see a quite good review by Jennen (1985)) is the use of a 'tangent approximation' to the first exit density of the form:

$$g_a(\lambda(t), t; x_0, t_0) = \frac{\lambda(t) - (t - t_0)\lambda'(t)}{(t - t_0)} f(\lambda(t), t; x_0, t_0)$$

For the case of the state of human health studied here Janssen and Skiadas (1995) proposed the following form:

$$g(t) = k\frac{|a|}{\sigma\sqrt{2\pi t^3}} \exp\left[-\frac{[a - \int_1^t \mu(s)ds]^2}{2\sigma^2 t}\right]$$

where k is a normalisation constant defined by the formula:

$$\int_0^\infty g(t)dt = 1$$

An investigation of the above form of $g(t)$ indicates that all the parameters except k are divided by σ or in other words when estimated they are estimated in units of σ. Thus we can let $\sigma = 1$ and use the expression for $\mu(t)$ to take the final form of $g(t)$ to be fitted on mortality table data. This form of $g(t)$ is the following:

$$g(t) = k\frac{|a|}{\sqrt{2\pi t^3}} \exp\left[-\frac{1}{2}\left(\frac{a + c}{\sqrt{t}} - (b_1 + a_1 t\sqrt{t} + a_2 t^2\sqrt{t} + a_3 t^3\sqrt{t})\right)^2\right]$$

77.2.3 Main parameters (Mean, Variance,...)

The parameters of the last formula are estimated by using a non-linear regression analysis technique. The mean and variance have no analytic expressions

b_1	$a_1 * 10^{-2}$	$a_2 * 10^{-4}$	$a_3 * 10^{-6}$	ka	MSE	r^2
4.990	-1.6459	2.6935	-1.7947	61.600	0.235	0.98

Table 77.1: Parameter Estimates MSE and R^2 for Greek Life-Table data (males, 1992)

for the general case. However, it is easy to obtain close approximations by performing arithmetic methods. The mean that provides the mean life time is estimated for each case during the applications that follow. The parameters of $g = g(t)$ are estimated by an iterative direct non-linear least squares method that leads to the minimization of the sum of the squared errors.

77.3 Application to the Greek Life-Table data (males, 1992) and Stochastic Simulation

The application was based on mortality tables of Greece for men for the year 1992. The data used for the fitting are those expressing the instantaneous rate of death $d(x)$ and are usually provided for 1.000.000 inhabitants. A non-linear regression analysis procedure was applied by using the formula for $g(t)$. The results for the cases studied are presented in Table 77.1.

The estimated model was used to generate a simulation procedure of various paths expressing the state of health of individuals and of the resulting density function for the hitting time. The simulation appears in Figure 77.1. The illustration of the original data and the estimated values is given in Graph C. The stochastic simulation is presented in Graph A. In Graph B the resulting density function for the hitting time after performing a large number of simulations appears. The total interval of the lifetime between 1 and 105 years is divided in 20 subintervals each of duration of 5 years and the interval over 105 years is also included. The results of the simulation study coincide with the real situation presented in Graph C. In Graph C the curves express the rate of death dx (real and estimated). The fitting was quite good as also is is verified by the high $R^2 = 0.98$ and the small Mean Squared Error $MSE = 2.351$. The Expected Life Time is found to be $ELT = 71.93$ years.

77.4 Conclusion

This paper presents a way of analysing mortality data using the concept of health state so that there now exists a dynamic study instead of a static one. With this application to Greek Life-Table data we show the high potentiality

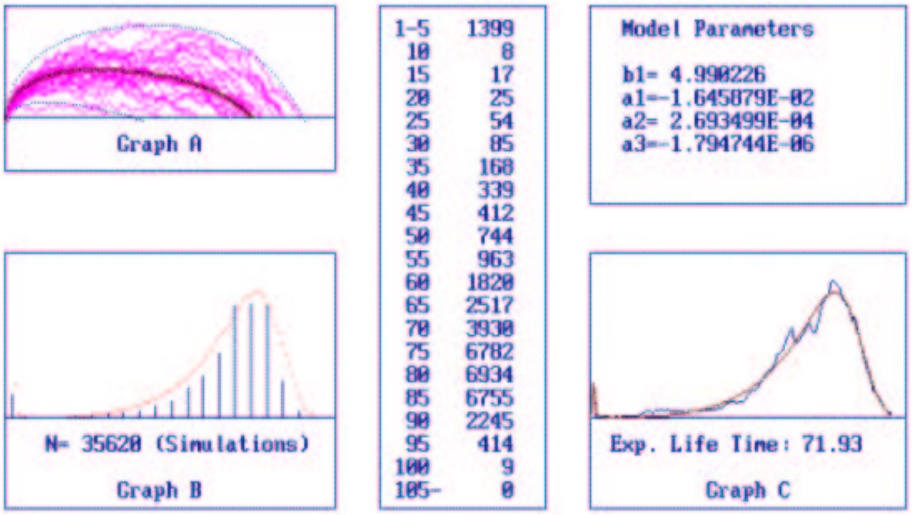


Figure 77.1: Stochastic simulation of Greek Life-Table data (males, 1992)

of such a dynamic stochastic approach, firstly to improve the present way of analysing mortality tables and secondly to provide a systematic tool for the comparison of several mortality tables in a time evolution or coming from different countries or regions. A more general study and a comparative study between countries will follow.

References

1. Buonocore,A., Nobile, A.G. and Ricciardi, L.M. (1987). A new integral equation for the evaluation of first-passage-time probability densities, *Advances in Applied Probability*, **19**, 784–800.
2. Giorno, V., Nobile, A.G. and Ricciardi, L.M. and Sato, S. (1989). On the evaluation of first-passage-time probability densities via non-singular integral equations, *Advances in Applied Probability*, **21**, 20–36.
3. Janssen, J. and Skiadas, C.H. (1995). Dynamic modelling of Life-Table data, *Applied Stochastic Models and Data Analysis*, vol. 11,**1**, 35–49.
4. Jennen, C. (1985). Second order approximations to the density mean and variance of Brownian first-exit times, *The Annals of Probability*, **13**, 126–144.
5. Siegert, A.J.F. (1951). On the first passage time probability problem, *Physical Review*, **81**, 617–623.

78

Using Independent Component Analysis of fMRI Time Series to Investigate Task-Related Activation

Mandy Sohr[1], **Waltraud Kahle**[2], **André Brechmann**[1]

[1] *Leibniz - Institute for Neurobiology*
Special Laboratory Non-Invasive Brain Imaging, Magdeburg, Germany
[2] *Otto-von-Guericke-University*
Institute for Mathematical Stochastics, Magdeburg, Germany

Abstract: The Independent Component Analysis (ICA) is a statistical and computational technique used to identify hidden factors of observed multivariate data. In this context the measured signals arise from functional Magnetic Resonance Imaging (fMRI) studies. FMRI is a non-invasive method used to study human brain functions by localizing activated brain areas and analyzing the intensity and time courses of neuronal activities. The time series obtained by fMRI studies are supposed to be linear mixtures of realizations of different stochastic processes, e.g. the neuronal responses to stimuli and the task performance of the subjects which are of special interest for neuroscientists. Additionally, there are processes related to heart beat or breathing of the subjects, motion artifacts, and noise caused by the tomograph. Since the exact temporal behavior of such signals is not always predicable, we used the ICA, a method without any hypothesis about the expected time courses, to extract signals possibly reflecting learning related processes from an auditory fMRI study with repeated measurements. First, we spatially localized the sources of neuronal activation related to the auditory task. Then, the temporal structure of the neuronal responses was analyzed with general time series statistics to extract and describe the signals statistically related to the task performance of the subjects. Performing the ICA to our data revealed activation cluster with associated time courses in auditory areas, attention-related areas, and somatosensory areas, among others. The most interesting finding was, that the time courses of these clusters showed different temporal behavior in the first two measurements compared to the last two measurements which might be explained by learning related effects. Classical methods for analyzing fMRI data like the General Linear Model (GLM) might fail to detect such activations.

Keywords: Functional magnetic resonance imaging, Independent Component Analysis, Time series

78.1 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive method used to study human brain functions by revealing which parts of the brain are involved in solving certain tasks. The fMRI methodology is based on the physical phenomenon that neural activity is expected to cause both a desoxygenation of blood and an increase of blood flow in vessels within the activated regions of the cortex. This effect is called 'blood oxygenation level dependent effect' (BOLD-effect) and is based on the different magnetic properties of deoxyhaemoglobin (blood with a low level of oxygen) and oxyhaemoglobin (blood with a high level of oxygen). Activated brain regions need more energy and thus consume more oxygen and glucose resulting in a change in the concentration of deoxyhaemoglobin and oxyhaemoglobin in this region (Ogawa *et al.*, 1990) and consequently changing the amplitude of the measured fMRI signal.

The time course of the fMRI signal is known as the Haemodynamic Response Function (HRF), which is the response to a temporary increase in neuronal activity. In typical fMRI measurements the stimuli are arranged in repeated blocks of about 30 seconds separated by resting blocks. In general, the fMRI signal is characterized by a delayed increase after the onset of stimulation, reaching a plateau level after about 6 seconds, and decreases slowly to baseline after the offset of stimulation in about 10 - 15 seconds.

During the fMRI measurement multiple slices of the brain are recorded, thereby each slice consists of a number of voxels (3D data points). Consequently, the fMRI images are composed of three-dimensional images at equidistant time points. For each voxel v_i, $i = 1, \ldots, N$, at the anatomical coordinate (v_x^i, v_y^i, v_z^i) in the human brain, we observe for each time point t $(t = 1, \ldots, T)$ the gray value $x_i(t)$, which can be represented as a function of the induced brain signal $f_t(v_i)$:

$$x_i(t) = f_t(v_i), \quad t = 1, \ldots, T. \tag{78.1.1}$$

How this signal behaves in the context of learning processes was investigated in this paper.

78.2 Independent Component Analysis

The Independent Component Analysis (ICA) is a method for blind signal separation (BSS) formed on the basis of assumed statistical independence of the source signals. This method transforms multidimensional data into components that are as statistically independent from each other as possible without

making any hypothesis about the theoretical time courses of the source signals. The observed signals $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))^T$ are assumed to be realizations of linear mixtures of stochastic processes. In the BSS problem, the underlying mixture model generates the observed signals $\mathbf{x}(t)$ from the M sources $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_M(t))^T$ $(M \leq N)$ by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{78.2.2}$$

where $\mathbf{A} = [a_{ij}]_{i=1,\ldots,N, j=1,\ldots,M}$ is the (unknown) mixing matrix. The source signals are assumed to be non-gaussian distributed. In the case of gaussian distributed signals, the signal separation problem is reduced to a Principal Component Analysis (PCA) using second-order statistics to estimate the source signals. The ICA algorithm aimed to find non-gaussian distributed source signals. For this purpose, the ICA uses either higher-order statistics like the kurtosis, which measures the gaussianity of a random variable, or the more theoretical mutual information. The mutual information is an information-theoretic function taking the whole dependency structure of the variables into account. Thus, finding a transform that minimizes the mutual information between the components is a natural way of estimating the ICA model (Comon, 1994).

Without knowing the source signals $\mathbf{s}(t)$ and the mixing matrix $\mathbf{A}$, ICA aims to recover the original sources from the observations $\mathbf{x}(t)$ by a linear invertible transformation

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t), \tag{78.2.3}$$

where $\mathbf{y}(t)$ are the estimates of the source signals $\mathbf{s}(t)$ $(\mathbf{y}(t) = \hat{\mathbf{s}}(t))$ and $\mathbf{W} = [w_{ji}]_{j=1,\ldots,M, i=1,\ldots,N}$ is the estimated unmixing matrix. $\mathbf{W}$ is determined such that the mutual information of the independent components $\mathbf{y}(t)$ is minimized. The matrix $\mathbf{W}$ is the pseudo-inverse of the mixing matrix, such that $\mathbf{AW} \approx \mathbf{I}$.

Before performing the ICA algorithm the observations were standardized. The standardization results in vectors $\mathbf{z}(t) = \mathbf{Z}\mathbf{x}(t)$ which all have mean zero and equal unit variances. The matrix $\mathbf{Z}$ is given by $\mathbf{Z} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$, where $\mathbf{\Lambda} = [\lambda(1), \ldots, \lambda(n)]$ is a diagonal matrix with the eigenvalues of the data covariance matrix $E\{\mathbf{x}(t)\mathbf{x}(t)^T\}$, and $\mathbf{U}$ is a matrix with the corresponding eigenvectors at its columns. At this step the dimension reduction can be performed by selecting only the M most interesting eigenvectors. In terms of the transformed vectors $\mathbf{z}(t)$, the model (78.2.2) becomes

$$\mathbf{z}(t) = \mathbf{Z}\mathbf{A}\mathbf{s}(t), \tag{78.2.4}$$

and the solution becomes

$$\mathbf{y}(t) = \mathbf{W}^T\mathbf{z}(t), \tag{78.2.5}$$

An algorithm that can then be used to extract the independent components is the FastICA algorithm proposed by Hyvarinen and Oja (1997). Since the mutual information is more a theoretical function, the kurtosis is used as a measure of gaussianity, which is defined by $kurt(y_i) = E\{y_i^4\} - 3[E\{y_i^2\}]^2$. Considering a linear combination $y = \mathbf{w}^T\mathbf{z}$ of a random vector $\mathbf{z}$, with $||\mathbf{w}|| = 1$, then $E\{y^2\} = 1$ and $kurt(y) = E\{y^4\} - 3$, whose gradient with respect to $\mathbf{w}$ is $4E\{\mathbf{z}(\mathbf{w}^T\mathbf{z})^3\}$. The FastICA algorithm maximizes the absolute value of the kurtosis, finds one of the columns $\mathbf{w}$ of the unmixing matrix $\mathbf{W}$, and thus identifies one independent component at a time using (78.2.5). The lth iteration of the algorithm is defined as

$$\begin{aligned} \mathbf{w}_l^* &= E\{\mathbf{z}(\mathbf{w}_{l-1}^T\mathbf{z})^3\} - 3\mathbf{w}_{l-1} \\ \mathbf{w}_l &= \mathbf{w}_l^*/||\mathbf{w}_l^*||. \end{aligned} \tag{78.2.6}$$

The ICA is an iterative method and to estimate more than one solution, and up to a maximum of M solutions, the algorithm must be run repeatedly.

78.3 ICA Applied to an fMRI Auditory Learning Study

The ICA was already successfully applied to fMRI data (see McKeown et al., 2003 for review). In this paper the ICA was applied to an auditory learning paradigm to characterize the fMRI signals, reflecting learning related neuronal processes.

Therefore, six subjects were scanned in a 3 Tesla scanner (Siemens Trio). Each subject was scanned five times over a period of five weeks. Frequency modulated (FM) tones were used as acoustic stimuli. They were arranged in stimulation blocks of 30 seconds. Each block consisted of 30 randomized FM tones, each 15 rising tones and 15 falling tones. The frequency range of the FM tones varied between 0.5 - 2 kHz. Six different durations were used; 300 ms, 350 ms 400 ms, 550 ms, 600 ms, and 650 ms (525 ms gap between two tones). The FM tones were presented at five different sound levels covering a range of 24 dB in steps of 6 dB all at a comfortable loudness. One experimental session consisted of 15 alternating stimulus and resting blocks. The whole fMRI data set consisted of 310 volumes (time points) by recording an fMRI image every 3 seconds (repetition time (TR) = 3000 ms; echo time (TE) = 30 ms; flip angle = 80°; field of view (FOV) = 192 mm; voxel size 64 × 64). 40 slices of 3 mm each (0.3 mm gap) covering the whole brain were recorded.

The task of the subjects during the measurement was a one-back working memory task. The subjects continuously had to compare the actual tone with the

tone presented one back in the sequence and had to indicate whether the two tones matched in direction (rising or falling) by button pressing. Since all subjects were naive according to FM tones, i.e. they never performed a discrimination task before this experiment, the task was quite difficult for the subjects at the beginning, but all subjects showed strong improvements indicated by their hit rates.

For each measurement and each subject an ICA with 30 independent components was performed. The independent components consisted of an activation map and associated time courses. According to their time-courses, the components can be classified as oscillatory functions, trend functions, noise functions, and some time courses possibly indicating neuronal processes. Considering the activation cluster of the latter time courses reflecting the neuronal processing, we found clusters in auditory areas, in many areas which are supposed to be involved in maintenance and attention processes and in areas which are involved in somato-sensory processes, caused by pressing a button to indicate targets. In a next step we investigated whether there are changes in the time course of these clusters over the five repeated measurements. A very interesting finding was, that the time courses of the cluster of the first two measurements showed the typical hemodynamic response function, i.e. the signal increases after stimulus onset, reaches a plateau, and decreases slowly after stimulus offset. But the time courses of the last two measurements showed a different behavior. The signal also increases after stimulus onset, mostly on a higher level than the signals in the first measurements, but the signal does not stay on the plateau it decreases immediately, such that at the end of stimulation the signal is already at baseline. This was often found for the auditory regions and areas involved in maintenance and attention processing and might be explained by adaptation, habituation or learning effects.

Classical methods for analyzing fMRI data like the General Linear Model (GLM) require a reference function describing the hypothetically expected time course of activated voxels. This reference function may correspond to the time course found in the first two measurements and is then correlated to the time course of each voxel to detect significantly activated voxels. Using this function also as reference function for the last two measurements may fail to detect possibly the most relevant activations.

Furthermore, simulation studies were performed using the signals reflecting different neuronal responses as source signals. These signals were linearly mixed with other signals reflecting oscillatory, linear or noise processes. The ICA was able to separate the signals reflecting different neuronal processing, i.e. signals of the first two measurement and signals of the last two measurements as well

as noise signals into independent components.

78.4 Summary

The ICA was applied to fMRI time series to detect signals reflecting learning related processes. The ICA was able to reveal regions of interest even if the time characteristics of these regions changed between repeated measurements. An advantage of ICA is, that it requires no prior information of hypothetical time courses of neuronal processes as it is needed in the General Linear Model (GLM), for instance. Consequently, the ICA seems to be an useful tool to investigate dynamic fMRI signals of repeated measurements.

References

1. Comon, P. (1994). Independent Component Analysis. A New Concept?, *Signal Processing*, **36**, 287–314

2. Hyvarinen, A., Oja, E. (1997). A Fast Fixed-Point Algorithm for Independent Component Analysis, *Neural Computation*, **9**, 1483–1492

3. McKeown, M.J., Hansen, L. K., Sejnowski, T. J. (2003). Independent Component Analysis of Functional MRI: What is Signal and What is Noise?, *Current Opinion In Neurobiology*, **13**(5), 620–629

4. Ogawa, S., Lee, T. M., Kay, A. R. and Tank, D. W. (1990). Brain Magnetic Resonance Imaging with Contrast Dependent on Blood Oxygenation, *Proceedings of the National Academy of Sciences of the United States of America*, **87**(24), 9868–9872

79

The Cell Proliferation and Apoptosis in the Presence of Amino Acids in Organotypic Culture of Tissues of Different Age

A.N. Zakutskii, N.I. Chalisova, A.I. Anisimova, S.V. Filippov

Pavlov Institute of Physiology, Russian Academy of Sciences **&** *St.Petersburg Institute of Bioregulation and Gerontology of the North-Western Branch of the Russian Academy of Medical Sciences*

79.1 Extended Abstract

Data confirming the concept that an organism possesses sufficiently independent regulatory systems-peptide and amino acid-have been accumulated to date [1, 4, 5]. For example, the studies of the parameters of specific and nonspecific resistance showed that lysine, arginine, glutamic and aspartic acids, and tryptophane exhibit different immunity and phagocytosis-stimulating and detoxicating properties [2]. The most adequate and convenient method for a rapid quantitative estimation of the direction of the effect of biologically active compounds is organotypic culturing of tissue fragments and analysis of the growth zone of explants. This is due to the fact that changes in the number of cells may serve as a criterion of primary integrated estimation of biological activity of substances, and a change in the number of cells itself may be the result of stimulation or inhibition of proliferation. Inhibition of proliferation due to apoptosis is studied by the methods of molecular biology, which allow detection of expression of proapoptotic proteins [6, 8]. Experiments were performed in organotypic culture with 1800 explants of fragments of the brain cortex, subcortical structures, cerebellum, spleen, and liver of one-day-old Wistar rats and 1900 explants of the same tissues of 21-day-old rats. The effect of each amino acid in each tissue was studied in 15–18 experimental transplants and in the same number of control explants. Tissue fragments, prepared under sterile conditions, were separated to smaller pieces which were placed into Petri dishes with collagen support. Nutrient medium consisted of 35and 5Equimolar solu-

tions of L-amino acids (Sigma, United States) were added to culture medium at an effective concentration of 0.05 ng/ml. Petri dishes were incubated for 3 days and then examined using a phase-contrast microscope equipped with a microtelemetric eyepiece. For each explant we determined the area index (AI), which was calculated as the ratio between the total area of explant (together with the zone of migrating cells) and the area of the central zone of explant and expressed in arbitrary units. The expression of the proapoptotic protein p53 was detected immunohistochemically [3, 7] using monoclonal antibodies to the p53 protein (1 : 75, Novocastra). Biotinylated antimouse and antirabbit immunoglobulins contained in the standard kit were used as secondary antibodies. Proteins were visualized using the complex of avidin with biotinylated horseradish peroxidase (ABC kit) with subsequent visualization of peroxidase with diaminobenzidine. All reagents were from Novocastra. Morphometric study was performed using the system of computer analysis of microscopic images. The results were statistically processed using the Statistika 5.0 software. The analysis of growth of explants of the brain cortex, subcortical structures, cerebellum, spleen, and liver of 1- and 21-day-old rats in the organotypic culture showed that the effects of different amino acids varied: the growth zone increased, decreased, or remained unchanged (in the last case, AI remained at the control level). The high-molecular-weight amino acids with low hydrophobicity (asparagine, lysine, arginine, and glutamic acid) had an inhibitory effect on the growth zone of explants of the brain cortex, spleen, and liver of one-day-old animals and an opposite, stimulatory effect on mature tissue of spleen and liver of 21-day-old rats. As mentioned earlier, such oppositely directed effects of the four amino acids, which depended on tissue maturity, were observed solely in tissues of the mesodermal and entodermal origin. In explants of subcortical structures, arginine and glutamic acid had a stimulatory effect on tissues of one-day-old rats, whereas the stimulatory effect of cerebellum of one-day-old rats was observed in the case of asparagine, lysine, arginine, glutamic acid, proline, valine, isoleucine, glycine, and cysteine. In mature tissues of subcortical structures of 21-day-old rats, the growth zone of explants increased in the presence of glutamine, arginine, cysteine, threonine, isoleucine, and tryptophane. In mature cerebellum, stimulatory effect was exerted only by cysteine and threonine. The stimulatory effect on the brain-cortex explants of 21-day-old rats was detected only for the low-molecular-weight amino acids exhibiting high hydrophobicity: AIs in the presence of aspartic acid, tyrosine, valine, threonine, methionine, and leucine increased by 42-57% compared to the control explants. In the case of the liver-tissue culture of 21-day-old rats, addition to nutrient medium of asparagine and arginine (0.05 ng/ml) statistically significantly (by 18-33%) increased the growth zone of explants compared to the control. Analysis of immunohistochemical preparations showed that the AI value was correlated with the expression of the p53 protein. When the

growth zone was depressed (and, therefore, the AIs values decreased), the expression of the proapoptotic protein p53 increased. To analyze the diversity of effects of amino acids, manifested in tissues of different genesis and different degree of maturity (five tissues of rats of two ages), it was necessary to calculate the frequencies of activity expression (stimulating either proliferation or apoptosis) of every amino acid. It can be seen that, in immature tissue, the highest frequencies of occurrence were characteristic of lysine, arginine, and glutamic acid; the lowest frequency, of histidine. An opposite picture was observed in mature tissues of 21-day-old rats: the highest frequency of occurrence was recorded for histidine and aspartic acid; the lowest frequency, for lysine and glutamic acid, which were the most active in immature tissue. Additionally, the mirror-type pattern was also observed in other cases; i.e., the amino acids that frequently occur in immature tissues rarely occur in mature tissues, and "vice versa." For example, glycine, asparagine, phenylalanine, and tryptophane occurred four times in immature tissues and two times in mature tissues; leucine occurred two times in immature tissues and five times in mature tissues; proline and isoleucine were detected two times in immature tissues and four times in mature tissues; and tyrosine and methionine were detected once in immature tissue and three times in mature tissues. The mirror-type proportion of the frequencies of activity of polar and some other amino acids apparently reflects the differences in the amino acid regulation of cell activity in immature and mature tissue.

References

1. Canete M., Juarranz A., Lopez-Nieva P., Alonso-Torcal C., Villanueva A., Stockert J.C. Fixation and permanent mouting of fluorescent probes after vital labelling of culture cells. Acta Histochem., 103 (2): 117-126. 2001.

2. Fratelli M., Gagliardini V., Galli G. et al. Autocrine interleukin-1 beta regulates both proliferation and apoptosis in EL4-6.1 thymoma cells. Blood. 85 (12): 3532-3537. 1995.

3. Gold R., Schmied M., Rothe G., Zischlier H., Breitschopf H., Wekerle H., Lassmann H. Detection of DNA fragmentation in apoptosis : application of in situ nick translation to cell culture systems and tissue sections. J. Histochem. Cytochem., 41: 1023-1030. 1993.

4. Itoh K., Hirohata S. The role of IL-10 in human B cell activation, proliferation and differentiation. J. Immunol. 154 (9): 4341-4350. 1995.

5. Lomo J., Blomhoff H.K., Beiske K. et al. TGF-beta 1 and cyclic AMP promote apoptosis in resting human B lymphocytes. J. Immunol. 154 (4): 1634-1643. 1995. [6] Mainou-Flowler T., Copplestone J.A., Prentice A.G. Effect of interleukins on the proliferation and survival of B cell chronic lymphocytic leukaemia cells. J. Clin. Pathol. 48 (5): 482-487. 1995.

6. Mekori Y.A., Oh C.K., Metcalfr D.D. The role of c-Kit and its ligand, stem cell factor, in mast cell apoptosis. Intern. Arch. Allergy Immunol. 107 (1-3): 136-138. 1995.

7. Panayiotidis P., Ganeshaguru K., Foroni L., Hoffbrand A.V. Expression and function of the FAS antigen in B chronic lymphocytic leukemia and hairy cell leukemia . Leukemia. 9 (7): 1227-1232. 1995.

8. Salas Vidal E., Lomeli H., Castro-Obregon S., Cuervo R., Escalante-Alcalde D., Covarrubias L. Reactive oxygen species participate in the control of mouse embryonic cell death. Exp. Cell Res. 238 (1): 136-147. 1998.

9. Trim N., Morgan S., Evans M., Issa R., Fine D., Afford S., Wilkins B., Iredale J. Hepatic stellate cells express the low affinity nerve growth factor receptor p75 and undergo apoptosis in response to nerve growth factor stimulation. Am. J. Pathol. 156 (4): 1235-1243. 2000.

80

Corrected Score Estimation in the Cox Regression Model with Misclassified Discrete Covariates

David M. Zucker†, Donna Spiegelman‡,

†*Department of Statistics, Hebrew University, Jerusalem, Israel*
‡*Departments of Epidemiology and Biostatistics, Harvard School of Public Health*

80.1 Introduction

We consider Cox survival regression with error-prone covariates. It is known that ignoring covariate error in regression analyses can lead to biased estimates of the regression coefficients. We focus here on discrete covariates subject to misclassification, which are of interest in many epidemiological studies. We also allow additional error-free covariates, which may be either discrete or continuous.

Three basic design setups are of interest: (1) the internal validation design, where the true covariate values are available on a subset of the main survival cohort, (2) the external validation design, where the measurement error distribution is estimated from data outside the main survival study, and (3) the replicate measurements design, where replicate surrogate covariate measurements are available, on either an internal or an external sample. Two types of models for the measurement error are of interest: structural models, where the true covariates are random variables, and functional models, where the true covariates are fixed values. Structural model methods generally involve estimation of some aspect of the distribution of the true covariate values; in functional model methods, this process is avoided.

The Cox model with covariate error has been examined in various settings. Our full paper gives a detailed review of the existing work. Much of this work focuses on the independent additive error model, under which the observed covariate value is equal to the true value plus a random error whose distribution is independent of the true value. For discrete covariates subject to misclassification, this model practically never holds, and so the methods built upon it do not apply. Other methods exist, but are subject to various limitations. There is a need for a convenient method for all three study designs that can handle gen-

eral measurement error structures, both functional and structural models, and time-dependent covariates. The aim of our work is to provide such a method for the case where the error-prone covariates are discrete, with misclassification of arbitrary form. Our method builds on a corrected score approach developed by Akazawa, Kinukawa, and Nakamura (1998) for generalized linear models. We begin by reviewing their work, and we then present our extension to the Cox model.

80.2 Review of the Corrected Score Technique

We work with a sample of n independent individuals. Associated with each individual i is a response variable T_i and a p-vector of covariates $\mathbf{X}_i$. The conditional density or mass function of T_i given $\mathbf{X}_i$ is denoted by $f(t|\mathbf{X}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a q-vector of unknown parameters, which includes regression coefficients and auxiliary parameters such as error variances. We have in mind mainly generalized linear models such as linear, logistic, and Poisson regression, but we present the theory in a general way. We denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. Extending Akazawa et al. (1998), we partition the vector $\mathbf{X}_i$ into $\mathbf{W}_i$ and $\mathbf{Z}_i$, where $\mathbf{W}_i$ is a p_1-vector of error-prone covariates and $\mathbf{Z}_i$ is a p_2-vector of error-free covariates. We denote the observed value of $\mathbf{W}_i$ by $\tilde{\mathbf{W}}_i$. The vector $\mathbf{W}_i$ is assumed to be discrete, with its possible values (each one a p_1-vector) denoted by $\mathbf{w}_1, \ldots, \mathbf{w}_K$. The range of values of $\tilde{\mathbf{W}}_i$ is assumed to be the same as that for $\mathbf{W}_i$. We denote by $k(i)$ the value of k such that $\tilde{\mathbf{W}}_i = \mathbf{w}_k$. The vector $\mathbf{Z}_i$ of error-free covariates is allowed to be either discrete or continuous. We denote $A_{kl}^{(i)} = \Pr(\tilde{\mathbf{W}}_i = \mathbf{w}_l | \mathbf{W}_i = \mathbf{w}_k, \mathbf{Z}_i, T_i)$, which defines a square matrix $\mathbf{A}^{(i)}$ of classification probabilities. We assume for now that $\mathbf{A}^{(i)}$ is known. We denote by $\mathbf{B}^{(i)}$ the matrix inverse of $\mathbf{A}^{(i)}$, which is assumed to exist. When individual i is a member of an internal validation sample, for the estimation of $\boldsymbol{\theta}$ we set $\tilde{\mathbf{W}}_i = \mathbf{W}_i$ and replace $\mathbf{A}^{(i)}$ by the identity matrix.

Define $\mathbf{u}(t, \mathbf{w}, \mathbf{z}, \boldsymbol{\theta}) = [\partial/\partial\boldsymbol{\theta}] \log f(t|\mathbf{w}, \mathbf{z}, \boldsymbol{\theta})$ and $\mathbf{u}_i(\boldsymbol{\theta}) = \mathbf{u}(T_i, \mathbf{W}_i, \mathbf{Z}_i, \boldsymbol{\theta})$. The classical normalized likelihood score function when there no covariate error is then given by $\mathbf{U}(\boldsymbol{\theta}) = n^{-1} \sum_i \mathbf{u}_i(\boldsymbol{\theta})$, and the maximum likelihood estimate (MLE) is obtained by solving the equation $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$. Under classical conditions, $E_{\ 0}[\mathbf{U}(\boldsymbol{\theta}_0)] = \mathbf{0}$ and the MLE is consistent and asymptotically normal. The idea of the corrected score approach is to find a function $\mathbf{u}^*(t, \tilde{\mathbf{w}}, \mathbf{z}, \boldsymbol{\theta})$ such that

$$E[\mathbf{u}^*(T_i, \tilde{\mathbf{W}}_i, \mathbf{Z}_i, \boldsymbol{\theta}) | \mathbf{W}_i, \mathbf{Z}_i, T_i] = \mathbf{u}(T_i, \mathbf{W}_i, \mathbf{Z}_i, \boldsymbol{\theta}). \tag{80.2.1}$$

Then, with $\mathbf{u}_i^*(\boldsymbol{\theta}) = \mathbf{u}^*(T_i, \tilde{\mathbf{W}}_i, \mathbf{Z}_i, \boldsymbol{\theta})$, we use the modified likelihood score function $\mathbf{U}^*(\boldsymbol{\theta}) = n^{-1} \sum_i \mathbf{u}_i^*(\boldsymbol{\theta})$ in place of $\mathbf{U}(\boldsymbol{\theta})$ as the basis for estimation. The estimation equation thus becomes $\mathbf{U}^*(\boldsymbol{\theta}) = \mathbf{0}$. In the case of discrete error-prone covariates, as shown by Akazawa et al. (1998), a function $\mathbf{u}^*$ satisfying

(80.2.1) is given by a simple formula:

$$\mathbf{u}_i^*(\boldsymbol{\theta}) = \sum_{l=1}^{K} B_{k(i)l}^{(i)} \mathbf{u}(T_i, \mathbf{w}_l, \mathbf{Z}_i, \boldsymbol{\theta}). \tag{80.2.2}$$

Let $\mathbf{J}_i(\boldsymbol{\theta})$ be the matrix with elements $J_{i,rs}(\boldsymbol{\theta}) = (\partial/\partial\theta_s)u_{i,r}(\theta)$ and let $\mathbf{J}_i^*(\boldsymbol{\theta})$ be defined correspondingly with $\mathbf{u}_i^*$ in place of $\mathbf{u}_i$.

Under the typical conditions assumed in generalized estimation equations (GEE) theory, the estimator $\hat{\boldsymbol{\theta}}$ will be consistent and asymptotically normal. The limiting covariance matrix $\mathbf{V}$ of $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ can be estimated using the sandwich estimator $\hat{\mathbf{V}} = \mathbf{D}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{H}(\hat{\boldsymbol{\theta}})\mathbf{D}(\hat{\boldsymbol{\theta}})^{-1}$, where $\mathbf{H}(\boldsymbol{\theta}) = n^{-1}\sum_i \mathbf{u}_i^*(\boldsymbol{\theta})\mathbf{u}_i^*(\boldsymbol{\theta})^T$ and $\mathbf{D}(\boldsymbol{\theta}) = -n^{-1}\sum_i \mathbf{J}_i^*(\boldsymbol{\theta})$.

The case where there are replicate measurements $\tilde{\mathbf{W}}_{ij}$ of $\tilde{\mathbf{W}}$ on the individuals in the main study can be handled in various ways. A simple approach is to redefine the quantity $\mathbf{u}_i^*(\boldsymbol{\theta})$ given in (80.2.2) by replacing $B_{k(i)l}^{(i)}$ with the mean of $B_{k(i,j)l}^{(i)}$ over the replicates for individual i, with $k(i,j)$ defined as the value of k such that $\tilde{\mathbf{W}}_{ij} = \mathbf{w}_k$. The development then proceeds as before.

80.3 Application to the Cox Survival Model

80.3.1 Setup

We now show how to apply the foregoing corrected score approach to the Cox model. Denote the survival time by T_i° and the censoring time by C_i. The observed survival data then consist of the observed follow-up time $T_i = \min(T_i^\circ, C_i)$ and the event indicator $\delta_i = I(T_i^\circ \le C_i)$. We let $Y_i(t) = I(T_i \ge t)$ denote the at-risk indicator. We assume the failure process and the censoring process are conditionally independent given the covariate process in the sense described by Kalbfleisch and Prentice (1980, Sec. 5.3.2).

The covariate structure is as described in the preceding section, except that the covariates are allowed to be time-dependent, so that we write $k(i,t)$ and $\mathbf{Z}_i(t)$. We assume that the measurement error process is "localized" in the sense that it depends only on the current true covariate value. More precisely, the assumption is that, conditional on the value of $\mathbf{X}_i(t)$, the value of $\tilde{\mathbf{W}}_i(t)$ is independent of the survival and censoring processes and of the values of $\mathbf{X}_i(s)$ for $s \ne t$. This assumption is plausible in many settings, e.g. when the main source of error is technical or laboratory error, or reading/coding error, as with diagnostic X-rays and dietary intake assessments. With no change in the theory, the classification probabilities $A_{kl}^{(i)}$ can be allowed to depend upon t. This extension permits accounting for improvements in measurement techniques over time. In addition, if internal validation data are available, this extension allows us to dispense with the localized error assumption.

In the proportional hazards model, the hazard function is taken to be of the form $\lambda(t|\mathbf{X}(t)) = \lambda_0(t)\psi(\mathbf{X}(t);\boldsymbol{\beta})$, with $\lambda_0(t)$ being a baseline hazard function of unspecified form. The function $\psi(\mathbf{x};\boldsymbol{\beta})$, which involves a p-vector $\boldsymbol{\beta}$ of unknown regression parameters which are to be estimated, represents the relative risk for an individual with covariate vector $\mathbf{x}$. The classical Cox model assumes $\psi(\mathbf{x};\boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T\mathbf{x}}$. We allow a general relative risk function satisfying $\psi(\mathbf{x};\mathbf{0}) = 1$, i.e. $\boldsymbol{\beta} = \mathbf{0}$ corresponds to no covariate effect. We let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$.

80.3.2 The Method

We now describe the method. Let $\psi_r'(\mathbf{x};\boldsymbol{\beta})$ denote the partial derivative of $\psi(\mathbf{x};\boldsymbol{\beta})$ with respect to β_r and define $\xi_r(\mathbf{x};\boldsymbol{\beta}) = \psi_r'(\mathbf{x};\boldsymbol{\beta})/\psi(\mathbf{x};\boldsymbol{\beta})$. Then the classical Cox partial likelihood score function in the case with no measurement error is given by

$$U_r(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\xi_r(\mathbf{X}_i(T_i);\boldsymbol{\beta}) - \frac{e_{1r}(T_i)}{e_0(T_i)}\right), \tag{80.3.3}$$

where

$$e_0(t) = \sum_{j=1}^{n} Y_j(t)\psi(\mathbf{X}_i(t);\boldsymbol{\beta}), \quad e_{1r}(t) = \sum_{j=1}^{n} Y_j(t)\psi_r'(\mathbf{X}_i(t);\boldsymbol{\beta}).$$

Now define

$$\psi_i^*(t,\boldsymbol{\beta}) = \sum_{l=1}^{K} B_{k(i,t)l}^{(i)}\psi(\mathbf{w}_l,\mathbf{Z}_i(t);\boldsymbol{\beta}), \quad \eta_{ir}(t,\boldsymbol{\beta}) = \sum_{l=1}^{K} B_{k(i,t)l}^{(i)}\psi_r'(\mathbf{w}_l,\mathbf{Z}_i(t);\boldsymbol{\beta}),$$

$$\xi_{ir}^*(t,\boldsymbol{\beta}) = \sum_{l=1}^{K} B_{k(i,t)l}^{(i)}\xi_r(\mathbf{w}_l,\mathbf{Z}_i(t);\boldsymbol{\beta}), \quad e_0^*(t) = \sum_{j=1}^{n} Y_j(t)\psi_j^*(t,\boldsymbol{\beta}),$$

$$e_{1r}^*(t) = \sum_{j=1}^{n} Y_j(t)\eta_{jr}(t,\boldsymbol{\beta}).$$

Then our proposed corrected score function is the following obvious analogue of (80.3.3):

$$U_r^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left(\xi_{ir}^*(T_i,\boldsymbol{\beta}) - \frac{e_{1r}^*(T_i)}{e_0^*(T_i)}\right). \tag{80.3.4}$$

As before, the proposed corrected score estimator is the solution to $\mathbf{U}^*(\boldsymbol{\beta}) = \mathbf{0}$, where $\mathbf{U}^*$ denotes the vector whose components are U_r^*.

Using an iterated expectation argument, under the localized error assumption, we can show that

$$E[Y_i(t)\psi_i^*(t,\boldsymbol{\beta})|\mathbf{X}_i(t)] = E[Y_i(t)\psi(\mathbf{X}_i(t);\boldsymbol{\beta})|\mathbf{X}_i(t)], \quad (80.3.5)$$

$$E[Y_i(t)\eta_{ir}^*(t,\boldsymbol{\beta})|\mathbf{X}_i(t)] = E[Y_i(t)\psi_r'(\mathbf{X}_i(t),\boldsymbol{\beta})|\mathbf{X}_i(t)], \quad (80.3.6)$$

$$E[Y_i(t)\xi_{ir}^*(t,\boldsymbol{\beta})|\mathbf{X}_i(t)] = E[Y_i(t)\xi_r(\mathbf{X}_i(t),\boldsymbol{\beta})|\mathbf{X}_i(t)]. \quad (80.3.7)$$

Thus, referring to the quantity in parentheses in (80.3.4), the first term and the numerator and denominator of the second term all have the correct expectation. It follows that $\mathbf{U}^*(\boldsymbol{\beta})$ is an asymptotically unbiased score function.

Accordingly, under standard conditions like those of Andersen and Gill (1982) and of Prentice and Self (1983), our corrected score estimator will be consistent and asymptotically normal. The asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ may be estimated by the sandwich formula $\hat{\mathbf{V}} = \mathbf{D}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{H}(\hat{\boldsymbol{\beta}})\mathbf{D}(\hat{\boldsymbol{\beta}})^{-1}$. Here $\mathbf{D}(\boldsymbol{\beta})$ is -1 times the matrix of derivatives of $\mathbf{U}^*(\boldsymbol{\beta})$ with respect to the components of $\boldsymbol{\beta}$ and $\mathbf{H}(\boldsymbol{\beta})$ is an empirical estimate of the covariance matrix of $\sqrt{n}\,\mathbf{U}^*(\boldsymbol{\beta})$.

The full paper gives the expressions for these matrices, an outline of the asymptotic argument, and an extension of the theory to the case where the classification matrix $\mathbf{A}^{(i)}$ is estimated. We also give results of a finite-sample simulation study under Weibull survival with a single binary covariate having known misclassification rates. The performance of the method described here was similar to that of related methods we have examined in previous work (Zucker and Spielgelman, 2004; Zucker, 2005). Specifically, our new estimator performed as well as or, in a few cases, better than the full Weibull maximum likelihood estimator. We also present simulation results for our method for the case where the misclassification probabilities are estimated from an external replicate measures study. Our method generally performed well in these simulations. We also illustrate the method on data from a study of the relationship between dietary calcium intake and distal colon cancer. The new estimator has a broader range of applicability than many other estimators proposed in the literature, including those described in our own earlier work, in that it can handle time-dependent covariates with an arbitrary misclassification structure.

References

1. Akazawa, K., Kinukawa, N., and Nakamura, T. (1998). A note on the corrected score function corrected for misclassification. *Journal of the Japan Statistical Society* **28**, 115–123.

2. Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.

3. Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

4. Kalbfleisch, J. D., and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley.

5. Prentice, R. L., and Self, S. G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *The Annals of Statistics* **11**, 804–812.

6. Zucker, D. M. (2005). A pseudo partial likelihood method for semiparametric survival regression with covariate errors. *Journal of the American Statistical Association*, **100**:1264-1277.

7. Zucker, D. M., and Spiegelman, D. (2004). Inference for the proportional hazards model with misclassified discrete-valued covariates. *Biometrics* **60**, 324–334.

Author (Participant) Index

Department of Mathematics and Statistics
Department of Biology
UNIVERSITY OF CYPRUS

es

european seminar

ΑΤΗΚ CYTA
Δίνει συνέχεια στην επικοινωνία!

SPSS

GENESIS Pharma (Cyprus) LTD

ΙΑΤΡΙΚΟ ΔΙΑΓΝΩΣΤΙΚΟ ΚΕΝΤΡΟ ΑΓΙΟΣ ΘΕΡΙΣΣΟΣ
MEDICAL DIAGNOSTIC CENTER AYIOS THERISSOS
M.R.I. - C.T. - U.S

Centre Culturel Francais en Chypre

ΔΗΜΟΣ ΓΕΡΜΑΣΟΓΕΙΑΣ 1994

CYPRUS TOURISM ORGANISATION

Kantzilaris Bookstores Ltd

Institute of Mathematical Statistics

ΝΕΑ Σ.Π.Ε. ΚΑ·Ι·ΜΑΚΛΙΟΥ

DEPARTMENT OF ANTIQUITIES
MINISTRY OF COMMUNICATIONS AND WORKS