Introduction

Dimitris Fouskakis

Department of Mathematics
School of Applied Mathematical and Physical Sciences
National Technical University of Athens

fouskakis@math.ntua.gr

Spring Semester

1. Density Estimation

$$\Omega \text{ (population)} \to X \text{ (characteristic, r.v.)} \sim f(x; \boldsymbol{\theta}) = ?$$
$$\mathbf{X} = (X_1, X_2 \ldots X_n) \text{ random sample} \to \hat{\boldsymbol{\theta}}$$

To estimate $\boldsymbol{\theta}$, we often need to make an assumption about the population distribution (or else about the model) f. How easy is to make an assumption?

1. If X is discrete and its description agrees with a Bernoulli, Binomial or Negative Binomial experiment.
2. If our aim is to use a sample statistic $T(\mathbf{X})$ which is a linear combination of $\sum_{i=1}^{n} X_i$ then for large n, independently of the choice of f, T follows approximately normal distribution because of the Central Limit Theorem:

$$\sum_{i=1}^{n} X_i \stackrel{\cdot}{\sim} \mathcal{N}(n\mu, n\sigma^2)\,,$$

where μ is the population mean and σ^2 the population variance.

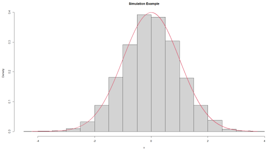
In other cases? $(X_1, X_2 \ldots X_n) \to \hat{f}$, i.e. estimation of f.

2. Stochastic Simulation

Any information we would like to know about a distribution, we can find it by simulating a large sample of values from it.

$$\text{Simulation} = \begin{cases} \text{reproduction of processes} \\ \text{mimicking the behavior of a model} \end{cases}$$

Using a computer → $(X_1, X_2 \ldots X_n) \sim f(x; \boldsymbol{\theta})$



$$\begin{aligned} \mu \quad &\rightarrow \quad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \\ \sigma^2 \quad &\rightarrow \quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \\ \mathbb{P}[X > 1] \quad &\rightarrow \quad \text{relative frequency} \\ \text{shape } f \quad &\rightarrow \quad \text{e.g. histogram} \end{aligned}$$

Strong Law of Large Numbers → Consistent estimators

3. Resampling Methods

1. JackKnife
2. Bootstrap
3. Cross-Validation

Precision of estimators? Bias of estimators ?

$$\bar{\mathbf{X}} \dot{\sim} \mathcal{N}(\mu, \sigma^2/n) \quad \text{from CLT}$$

Median?

$$\longrightarrow \quad \text{Resampling methods} = \begin{cases} \text{Bootstrap} \\ \text{Jackknife} \end{cases}$$

4. Model Selection

How "good" is our model? If we have two models, which one to choose?

Answer: use Cross-Validation.

Idea: Split data into a modelling and a validation sub-sample. Fit competing models into the modelling sub-sample and then compare their predictive accuracy into the modelling sub-sample.

5. Expectation - Maximization Algorithm

$$\Omega \text{ (population)} \to X \text{ (characteristic, r.v.)} \sim f(x; \boldsymbol{\theta})$$
$$\mathbf{X} = (X_1, X_2 \ldots X_n) \text{ random sample} \to \hat{\boldsymbol{\theta}}$$

To estimate $\boldsymbol{\theta}$, we often maximizing the Likelihood Function. Can we always do this analytically (using derivatives)?

No! For example if we have the gamma distribution with both parameters unknown.

Solution:

1. Use a Numerical Analysis Algorithm, e.g. Newton Raphson.
2. Use a Statistical Method, that takes into account the statistical model → Expectation-Maximization (EM).

6. Stochastic Optimization

In Statistics we often come up against maximization/minimization problems. For example in order to find the maximum likelihood estimator (MLE) → function maximization → there are functions that cannot be maximized analytically.

$$\text{e.g. } g(x) = \tfrac{\log x}{1+x} \rightarrow \text{no analytical solution}$$

Apart from MLE problems, statisticians come up with many other optimization problems:

i. Bayesian Decision Theory → cost minimization
ii. Solving non-linear least squares problems
iii. Choosing an appropriate model (e.g. variable selection)

The problem is always the same: maximizing (or minimizing) a real function g w.r.t a p-dimensional vector $\mathbf{x}$.
e.g. MLE $g \rightarrow log(L)$, L : likelihood function and $\mathbf{x} \rightarrow \boldsymbol{\theta} = (\theta_1, \ldots \theta_p)$: parameter vector

$$\hat{\boldsymbol{\theta}} : g'(\boldsymbol{\theta}) = 0 \iff \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \ldots \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_p} \right) = (0, \ldots 0)$$

It might not be possible to find an analytical solution.
linear equations → SIMPLEX. Non-linear equations?

6. Stochastic Optimization (cont'd)

For smooth, non-linear, differentiable functions: numerical solution (e.g. bisection method, Newton-Raphson method, secant method, Gauss-Newton method)

Consider now that we would like to maximize

$$f(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = (\theta_1, \dots \theta_p) \in \Theta \quad (\Theta : \text{discrete with } N \text{ elements})$$

[Combinatorial optimization (stochastic) → heuristic techniques (gradual improvement & local neighborhood)]

Every $\boldsymbol{\theta} \in \Theta$ is a candidate solution. Let $f_{\max}$ be the maximum and $\mathcal{M} = \{\boldsymbol{\theta} \in \Theta : f(\boldsymbol{\theta}) = f_{\max}\}$ (might contain more than one elements).

If N is large and there are several local maxima, finding the elements of $\mathcal{M}$ is hard (e.g. travelling salesman problem).

In general, p - objects combined in a large number (N) and every choice is a possible solution, e.g. travelling salesman problem $N = p!$

6. Stochastic Optimization (cont'd)

1. Genetic Algorithm
2. Simulated Annealing
3. Tabu Search

Necessity of using heuristic methods

It is not possible to use algorithms which is certain they are going to find the global maximum but in non-practical time. In contrast, we are working with algorithms which is possible to find global/local or nearby maxima in specific time, i.e. heuristics:

i. gradual improvement
ii. local neighborhood

7. Variable selection in linear regression

$$Y = b_0 + b_1 X_1 + \ldots + b_p X_p + \epsilon\,, \quad \epsilon \sim \mathcal{N}(0.\sigma^2)$$

Given a sample of size n, which of $X_1, \ldots X_p$ to use? $\begin{cases} \text{parsimony} \\ \text{goodness of fit} \end{cases}$

$$\text{Models}: \quad \boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)\,, \quad \gamma_i = \begin{cases} 1 & (i \text{ variable included}) \\ 0 & (i \text{ variable not included}) \end{cases}$$

The space of all possible models:

$$\mathcal{M} : 2^p elements$$

If p is large then the number of models is huge.

7. Variable selection in linear regression (cont.)

- AIC, BIC minimization
- If $p > n$? → shrinkage methods (e.g. Ridge, Lasso)

Minimization of AIC or BIC: model simplicity vs predictive accuracy

$$AIC = n \ln \frac{RSS}{n} + 2(S+2) = -2 \ln f(y|\hat{\theta}) + 2k \,,$$

n: sample size, RSS: residual sum of squares, S: $\#$ of parameters

$$BIC = -2 \ln f(y|\hat{\theta}) + k \ln n$$