Bayesian Variable Selection in Generalised Linear Models using a Combination of Stochastic Optimisation Methods

Dimitris Fouskakis,

Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Athens, Greece; e-mail: fouskakis@math.ntua.gr.

Presentation is available at:

www.math.ntua.gr/~fouskakis/Presentations/Southampton/Presentation_Southampton.pdf.

Reference:

Synopsis

1. Introduction
2. Bayesian Variable Selection
3. Prior Specification
4. Bayesian Model Averaging (BMA)
5. Stochastic Optimization Algorithms as a Model Search Tool
6. Combination of Stochastic Optimization Methods
7. Experimental Results
8. Discussion
1 Introduction

What is Model Selection?

• Evaluation of performance of scientific scenarios
• Selection of the “best”.

“Best” Model?

• The “best” performed model is totally subjective.
  – It may not be possible to find a single model capturing the preferences of all relevant stakeholders in the visited problem.
• Different procedures (or scientists) support different scientific theories.

All Models are wrong, but some are useful: George, E.P. Box

Main Principles: Goodness of fit vs. Parsimony
Available Methods

Classical Model Selection

- Significance Tests
- Stepwise Model Search (Forward Strategy, Backward Elimination)

Bayesian Model Selection

- Posterior Odds and Model Probabilities - BMA - BIC
- Utility Measures
- Predictive Measures
- Deviance Information Criteria

Information Criteria

- AIC
- BIC
- DIC
Stepwise Procedures

- Large datasets ⇒ small p-values even if the hypothesized model is plausible.
- Stepwise methods are sequential application of simple significance tests ⇒ exact significance level cannot be calculated.
- The maximum $F$-to-enter statistic is not even remotely like an F distribution.
- The selection of a single model ignores model uncertainty.
- Compare only nested models.
- Different procedures or starting from different models ⇒ different selected models.
- Prior knowledge is ignored, wasteful of information when good information exists.
1. The **Bayesian approach** is fundamentally probabilistic. A joint probability distribution is used to describe the relationships between all the unknown and the data. This joint distribution is obtained by introducing **prior distributions** on all the unknowns. Inference is then based on the conditional probability distribution of the unknowns given the observed data, the **posterior distribution**.

2. **Model Uncertainty**: A large number of different models are under consideration for the data. The joint distribution is obtained by introducing a prior distribution on all the unknowns, the **parameters** of each model and the **models** themselves. Conditioning on the data then induces a posterior distribution of model uncertainty that can be used for model selection and other inference and decision problems.

3. **Variable Selection**: A special case of the model selection problem where each model corresponds to a distinct subset of the available explanatory variables.

4. **Two main challenges**: Specification of the prior distributions and the calculation of the posterior.
2 Bayesian Variable Selection

Within the Bayesian framework the identification of the “best set of predictors” between the $\mathcal{M} = \{0, 1\}^p$ competitors is equivalent (assuming a zero-one loss function) to find the model $m_k$ with the highest posterior model probability, defined as

$$f(m_k | y) = \frac{f(y | m_k) f(m_k)}{\sum_{m_\ell \in \mathcal{M}} f(y | m_\ell) f(m_\ell)},$$

(1)

where $f(y | m_k)$ is the marginal likelihood (or prior predictive or model evidence) under model $m_k$ and $f(m_k)$ is the prior probability of model $m_k$. The marginal likelihood function in the above calculation can be further expanded to include the effect of the model parameters:

$$f(y | m_k) = \int f(y | \theta_k, m_k) f(\theta_k | m_k) d\theta_k,$$

(2)

where $f(y | \theta_k, m_k)$ is the likelihood under model $m_k$ with parameters $\theta_k$ and $f(\theta_k | m_k)$ is the prior distribution of model parameters given model $m_k$. 

Based on the above posterior model probabilities, pairwise comparisons of any two models, \( m_k \) and \( m_\ell \), is given by the **Posterior Odds** (PO)

\[
P_{O_{m_k,m_\ell}} \equiv \frac{f(m_k|y)}{f(m_\ell|y)} = \frac{f(y|m_k)}{f(y|m_\ell)} \times \frac{f(m_k)}{f(m_\ell)} = B_{m_k,m_\ell} \times O_{m_k,m_\ell} \tag{3}
\]

which is a function of the **Bayes Factor** \( B_{m_k,m_\ell} \) and the **Prior Odds** \( O_{m_k,m_\ell} \).

The posterior model probability (1) can be then expressed entirely in terms of Bayes Factors and Prior Odds as

\[
f(m|y) = \left( \sum_{m_\ell \in \mathcal{M}} P_{O_{m_\ell,m}} \right)^{-1} = \left[ \sum_{m_\ell \in \mathcal{M}} B_{m_\ell,m} \times O_{m_\ell,m} \right]^{-1}. \tag{4}
\]
Bayes Factor Interpretation - Kass and Raftery (1995, JASA)

\[ H_0 : m_0 \text{ vs. } H_1 : m_1 \]

<table>
<thead>
<tr>
<th>( \log(B_{m_1,m_0}) )</th>
<th>( B_{m_1,m_0} )</th>
<th>Evidence against ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>1-3</td>
<td>Negligible</td>
</tr>
<tr>
<td>1-3</td>
<td>3-20</td>
<td>Positive</td>
</tr>
<tr>
<td>3-5</td>
<td>20-150</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt; 5</td>
<td>&gt; 150</td>
<td>Very Strong</td>
</tr>
</tbody>
</table>
Under the decision theoretic framework let $\alpha$ represent the action of selecting $m_k$ and suppose that $\alpha$ is evaluated by a utility function $u(\alpha, \Delta)$, where $\Delta$ is some unknown quantity of interest, e.g. a future observation. The optimal selection is that $\alpha$ that maximizes the expected utility

$$
\int u(\alpha, \Delta) p(\Delta | y) d\Delta,
$$

where the predictive distribution of $\Delta$ given $y$

$$
p(\Delta | y) = \sum_k p(\Delta | m_k, y) p(m_k | y)
$$

is a posterior weighted mixture of the conditional predictive distributions

$$
p(\Delta | m_k, y) = \int p(\Delta | \theta_k, m_k) p(\theta_k | m_k, y) d\theta_k.
$$
1. If $\Delta$ identifies one of the competing models as the “true state of nature” and $u(\alpha, \Delta)$ is 0 or 1 according to whether a correct selection has been made, then selection of the highest posterior probability model will maximize expected utility.

2. **Logarithmic Score function:** $u(\alpha, \Delta) = \log p(\Delta|m_k, y)$ and $\Delta$ is a future observation. The best then selection is that $\alpha$ that maximizes the posterior weighted logarithmic divergence (San Martini & Spezzaferri, 1984, JRSS B)

   \[
   \sum_{\ell} p(m_\ell|y) \int p(\Delta|m_\ell, y) \log \frac{p(\Delta|m_\ell, y)}{p(\Delta|m_k, y)} dy
   \]  

   (8)

3. If the goal is strictly prediction and not model selection then expected logarithmic utility is maximized by using the posterior weighted mixture $p(\Delta|y)$. Under square error loss the best prediction of $\Delta$ is the overall posterior mean

   \[
   E(\Delta|y) = \sum_k E(\Delta|m_k, y) p(m_k|y).
   \]  

   (9)
1. Closed form expression of the marginal likelihood

- BIC approximation (Schwarz, 1978, Ann. Statist)
- Laplace approximation (Raftery, 1996, Biometrika)
- Monte Carlo Estimates
  - Naive (Sampling from the Prior)
  - Harmonic Mean (Sampling from the Posterior; Kass & Raftery, 1995, JASA)
  - Importance Sampling (Newton & Raftery, 1994, JRSS B)
- MCMC Estimates
  - Chib’s Estimator (Chib, 1995, JASA)
2. Exploration of the model space & estimation of the posterior

- Stochastic Optimization Algorithms (e.g. Fouskakis & Draper, 2008, JASA)
  - Simulated Annealing
  - Genetic Algorithm
  - Tabu Search

- MCMC Algorithms (*the first 3 only for variable selection problems*)
  - Gibbs Variable Selection (Dellaportas et al., 2002, Stat. & Comp)
  - Stochastic Search Variable Selection (George & McCulloch, 1993, JASA)
  - Kuo-Mallick Sampler (Kuo & Mallick, 1998, Sankya B)
  - Carlin-Chib Method (Carlin & Chib, 1995, JRSS B)
  - Reversible Jump MCMC (Green, 1995, Biometrika)
One of the most popular approximation techniques is the Bayesian information criterion (BIC) (Schwarz, 1978, Ann. Statist). The BIC value for any model $m$ is given by:

$$BIC(m) \equiv D(\hat{\theta}_m, m) + dm \log n,$$

where $D(\hat{\theta}_m, m) = -2\log(y|\hat{\theta}_m, m)$ is the usual deviance measure of model $m$ evaluated under the maximum likelihood estimates $\hat{\theta}_m$ of the model $m$ parameters and $dm$ is the dimension of $\theta_m$. From the above definition it is evident that the BIC is a penalized deviance measure with penalty equal to $\log n$ for each parameter estimated by the model. For any two models $m_k$ and $m_\ell$, Schwarz showed that

$$\frac{S_{m_k,m_\ell} - \log(B_{m_k,m_\ell})}{\log(B_{m_k,m_\ell})} \to 0 \text{ as } n \to +\infty,$$

where $S_{m_k,m_\ell} = -0.5\{BIC(m_k) - BIC(m_\ell)\}$. Therefore from (11) we obtain:

$$-2\log(B_{m_k,m_\ell}) \approx BIC(m_k) - BIC(m_\ell).$$
A more accurate approximation of the posterior distribution of a model $m$ can be derived using Laplace approximation (Raftery, 1996, Biometrika):

$$-2 \log f(m|y) = -2 \log f(y|\tilde{\theta}_m, m) - 2 \log f(\tilde{\theta}_m|m) - d_m \log(2\pi)$$

$$- \log |\Psi_m| - 2 \log f(m) + O(n^{-1}),$$

(13)

where $\tilde{\theta}_m$ is the posterior mode of $f(\theta_m|y, m)$, $d_m$ is the dimension of the model $m$ and $\Psi_m$ is minus the inverse of the Hessian matrix of $h(\theta_m) = \log f(y|\theta_m, m) + \log f(\theta_m|m)$ evaluated at the posterior mode $\tilde{\theta}_m$.

**Laplace - Metropolis Estimator** (Lewis and Raftery, 1997, JASA): Use the posterior mean or median (estimated from an MCMC) instead of the posterior mode and estimate the posterior variance from an MCMC output. The posterior should be symmetric (or close).
Alternative Representation

- The set of models $\mathcal{M}$ under consideration, can be represented by a vector of binary indicators $\gamma = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$, denoting which explanatory variables are present in the linear predictor.

- The $\gamma_j$ takes the value 1 if variable $j$ is included in the model and 0 otherwise.

- There is a one-to-one transformation connecting the usual model indicator $m$ with $\gamma$. One way to transform $\gamma$ to a unique model indicator $m$, is to consider $\gamma$ as a number in the binary numerical system and transform it to the corresponding number in the decimal numerical system.
3 Prior Specification

- Usually a uniform prior on model space is used, i.e.,

\[
f(m) \propto I [m \in \mathcal{M}].
\]  

(14)

When using the variable selection indicators \( \gamma \) this prior is equivalent to specifying independent Bernoulli prior distributions with inclusion probability equal to 0.5, i.e. \( \gamma_j \sim Bernoulli(0.5) \) for all \( j = 1, \ldots, p \). The above prior can be considered as informative since it puts more weight on models of size close to \( p/2 \).

- As a generalization of the above prior is to use a prior of the form:

\[
f(\gamma) = \prod_{i=1}^{p} \omega_i^{\gamma_i} (1 - \omega_i)^{1-\gamma_i}, \quad \omega_i \in [0, 1]
\]  

(15)

This prior can be used to put increased weight on parsimonious models by setting the \( \omega_i \) small.
Finally, following the work by Ley and Steel (2009, Journal of Applied Econometrics) we can use a **beta-binomial hierarchical prior**. If we denote by $d\gamma$ the model size then this prior has the following form

$$d\gamma \equiv \sum_{j=1}^{p} \gamma_j \sim \text{Bin}(p, \theta) \text{ and } \theta \sim \text{Beta}(\alpha, \beta). \quad (16)$$

The above prior depends on two hyperparameters, $(\alpha, \beta)$. For fix $\alpha = 1$ the above prior could be elicited in terms of the prior mean model size, $m$. The choice of $m$ will then determine $\beta$ through the formula $\beta = (p - m)/m$. By setting $m = p/2$ will get $\beta = 1$ and therefore we obtain a discrete uniform prior for model size. We can easily show that the prior probability for any model $\gamma_\ell$ of size $k$, using the prior (16), with $\alpha = 1$, is given by:

$$P(\gamma_\ell) = \frac{\Gamma\left(1 + \frac{p-m}{m}\right)}{\Gamma\left(\frac{p-m}{m}\right)} \frac{\Gamma(1 + k)\Gamma\left(\frac{p-m}{m} + p - k\right)}{\Gamma(1 + \frac{p-m}{m} + p)}. \quad (17)$$
Normal Models

Let $\theta_m = (\beta_m, \sigma^2)$ denote the unknown parameter vector for model $m$.

**Normal Inverse Gamma (NIG) Conjugate Prior**

$$\beta_m|m, \sigma^2 \sim N(\mu_{\beta_m}, g^2 V_m \sigma^2)$$

$$\sigma^2|m \sim IG(a, b)$$

- Marginal Likelihood is analytically tractable.
- Main issue: specification of $g^2 V_m$. 
Normal Models

**Zellner's g-prior** (Zellner, 1986):

\[ \text{NIG with } \mu_{\beta_m} = 0 \text{ and } V_m = (X_m^T X_m)^{-1}. \]

- For \( g = n \) ⇒ **unit information prior** (Kass & Wasserman, 1995, JASA)
- Can be extended to GLMs.
Prior on model parameters

Empirical Unit Information Prior:

\[ \theta_m|m \sim N(\hat{\theta}_m, n[I(\hat{\theta}_m)]^{-1}) , \]

where \( \hat{\theta}_m \) is the MLE and \( I(\hat{\theta}_m) \) the observed Fisher information matrix.

- Information equal to one data point.
- Uses data but minimally. It is still empirical.
- Behavior approximately equal to BIC.
Prior on model parameters


\[ f(\theta_m|m) \propto f(y^*|\theta_m, m)^{1/c^2}. \]

- \( y^* \): imaginary data.
- \( c^2 \): controls the weight given to imaginary data.
- \( c^2 = n \): accounts for one data point (Unit information prior).
Prior on model parameters

Zellner and Siow Prior (Zellner & Siow, 1980, Bayesian Statistics 1):

\[ \theta_m | m \sim Cauchy. \]

- Mean and Variance similar to the Zellner’s g-prior.

Mixtures of g-priors (Liang et al., 2008, JASA):

- Use Zellner’s g-prior.
- Put a prior on g.
- \[ \frac{g}{1+g} \sim Beta(1, \frac{\alpha}{2} - 1). \]
- \( 2 \leq \alpha \leq 4. \)
Prior on model parameters

Some comments:

- Normal Priors $\Rightarrow$ Ridge regression type of shrinkage.
- Double exponential priors $\Rightarrow$ LASSO regression type of shrinkage.
- Multivariate structure is important, use independent priors if there is no correlation between covariates.
- Ghosh and Clyde (2011, JASA) have worked on the idea of augmented the non orthogonal design matrix to an orthogonal one. Then independent priors can be used.
Lindley-Bartlett Paradox

- Diffuse reference priors (with large variance) cannot be used for the model parameters.
- When comparing two nested models with diffuse priors, the Bayes factor systematically favours the smaller model, regardless of the information in the data.
Variants of Bayes Factors

What if we use **improper priors** (e.g. default priors) for **objectivity**, e.g. Jeffrey’s prior, i.e. if $I(\theta_m)$ is the Fisher’s information matrix then $f(\theta_m|m) \propto \sqrt{\text{det}I(\theta_m)}$?

**Improper priors** for the model parameters, cannot be used since normalising constants involved in the Bayes factors (BFs) cannot be determined.

**Solutions:**

- Intrinsic Bayes factor (IBF), Berger and Pericchi (1996, JASA).
Intrinsic Bayes factor

- Use part of the data as a training sample. Then use the resulting posterior as the prior for the remaining data.
- The resulting Bayes factor is the **Intrinsic Bayes factor (IBF)** (Berger and Pericchi, 1996, JASA).
- To reduce the dependence on the specific training sample, take the average of the IBFs over all possible training samples ⇒ **AIBF**.
- You can start with an **improper baseline prior**.
- **Intrinsic Priors**: Priors that would yield Bayes factors that are approximately equal to IBS’s in an asymptotic case.
Expected-posterior priors (EPPs) (Pèrez and Berger, 2002, Biometrika) are defined as the posterior distribution of a parameter vector for the model under consideration, averaged over all possible imaginary samples $\mathbf{y}^*$ coming from a “suitable” predictive distribution $m^*(\mathbf{y}^*)$. Hence the EPP for the parameters of any model $m_\ell \in \mathcal{M}$ is

$$
\pi^E_\ell (\theta_\ell) = \int f(\theta_\ell | \mathbf{y}^*, m_\ell) m^*(\mathbf{y}^*) \, d\mathbf{y}^*
$$

**Base-model Approach:** Select a “reference” model $m_0$ for the training sample and define $m^*(\mathbf{y}^*) = m^N_0 (\mathbf{y}^*) \equiv f(\mathbf{y}^* | m_0)$ to be the prior predictive distribution, evaluated at $\mathbf{y}^*$, for the reference model $m_0$ under the baseline prior $\pi^N_0 (\theta_0)$. The reference model should be at least as simple as the other competing models, and therefore a reasonable choice is to take $m_0$ to be the constant model (with no predictors). **Impropriety of priors causes no indeterminacy.**
4 Bayesian Model Averaging (BMA)

- When working with Bayes factors, the decision space involves the choice of a model that is then used for inference or prediction.
- If the chosen model is only one of many possibilities, we run the risk that model uncertainty will be ignored.
- It makes more sense to look at the panoply of (some “good”, $K$ in total) models and the inferences or predictions they would give.
- Suppose, for example that the quantity of interest is $\Delta$ (a future observation for example). Then report:

$$f(\Delta|\mathbf{y}) = \sum_{k=1}^{K} f(\Delta|\mathbf{y}, m_k) f(m_k|\mathbf{y}).$$

This is a weighted average of the posterior probabilities of $\Delta$ under each model, where the weights are given by the posterior probabilities of the models in question. Raftery *et al.* (1997, JASA) called this approach *Bayesian Model Averaging* (BMA).
5 Stochastic Optimization Algorithms as a Model Search Tool

- The number of models under consideration is equal to $2^p$ and therefore when $p$ is even moderately large (e.g. $p = 50$), this number grows tremendously (e.g. $1.2 \times 10^{15}$). As a result, visiting every possible competing model becomes infeasible.

- This motivates the need for global optimization methods, for example stochastic optimization techniques.

- Alternatively, popular MCMC methods can be used (e.g. RJMCMC or $MC^3$), but they may suffer from poor mixing in high dimensional spaces, and as a result they may be unable to explore the full support of the posterior distribution.
Simulated annealing (SA) (Kirkpatrick *et al.* , 1983, Science) is a method closely related to Metropolis sampling except that instead of trying to sample from the function of interest you are trying to maximise it. Let’s suppose that our goal is to approximate the maximum of a criterion/objective function $\varphi: \mathcal{M} \rightarrow \mathbb{R}$.

- Choose a random initial model $m_0$. Set $m = m_0$. Set the initial temperature $T_0$;
- Repeat:
  - Choose at random a neighboring model $m'$;
  - If $\varphi(m') > \varphi(m)$ then set $m = m'$;
  - else set $m = m'$ with probability
    \[ \exp \left\{ \frac{\varphi(m') - \varphi(m)}{T} \right\} ; \] (18)
  - let $T$ decrease according to some schedule (e.g. Geometric or Reciprocal);
- Until the final temperature $T_f$ is reached.
Genetic Algorithm

A method first proposed by Holland (1975, University of Michigan Press). Its goal is to maximize an objective function $\varphi(\gamma)$ of binary strings of length $p$.

- Randomly generate an even number $n$ of models and compute their objective function (fitness).
- Select $n$ models for the next population with replacement from the initial population, with probability proportional to the fitness of every model.
- Consider the models of the previous population in pairs and perform the crossover operation with probability $p_c$: Take every pair and if crossover occurs, generate an integer $k$ from $U(1, p - 1)$, where $p$ is the number of variables, and the last $(p - k)$ elements of each model are exchanged to create 2 new models.
- Consider the previous population and perform the mutation operation for each variable of each model with probability $p_m$: If mutation occurs then flip the value of the variable from 0 to 1 or vice versa.
- Store best result and continue the algorithm for a specified number of iterations.
Genetic Algorithm (continued)

Modifications:

- Different crossover rules can be also applied, e.g. the uniform or the highly uniform crossover (see for example Fouskakis and Draper 2008, JASA).

- Elitist strategies can be also applied during the crossover step; instead of copying the offspring to the new population, the two fittest configurations between the two parents and the two offspring are copied to the new population.
Tabu search (TS) is a “higher level” heuristic procedure for solving large optimisation problems, proposed by Glover (1989, ORSA Journal of Computing). TS has three phases: preliminary search, intensification, and diversification.

During preliminary search TS is similar to some other optimisation methods in that, whatever model $m$ in the input space you are currently at, you evaluate the criterion function at all the neighbours $\mathcal{N}$ of $m$ and find the new model $m'$ that is best in $\mathcal{N}$, but TS differs from many other methods in that you move to $m'$ even if it is worse than $m$.

Repeating this idea creates the possibility of endlessly cycling back and forth between $m$ and $m'$; to avoid this TS uses the idea of a tabu list of forbidden moves, so that (e.g.) once the move $m \rightarrow m'$ has been made the reverse move $m' \rightarrow m$ is forbidden for at least the next $s$ moves.
One potential problem with the tabu list is that **it may forbid certain relevant or interesting moves**, for example those that lead to a better $x$ than the best one found so far.

Consequently, an **aspiration criterion** is introduced to allow moves that would otherwise be tabu to be chosen anyway, if they are judged to be worthwhile.

In the second (**intensification**) part of the search, you (a) start with the best solution found so far (which is always stored throughout the entire algorithm), (b) clear the tabu list, and (c) proceed as in the preliminary search for a specified number of moves. (This step can be **restarted randomly** a given number of times.)

Finally, in the **diversification** phase, you again clear the tabu list, and set the $s$ most frequent moves of the run so far to be tabu.

Then you choose a random $m$ to move to and **proceed as in the preliminary search phase** for a specified number of iterations.
6 Combination of Stochastic Optimization Methods

Part A (SATS algorithm)

We add features of TS in SA and use a variety of neighborhood structures.

• With probability 0.4 the algorithm proposes all 1-bit flip moves sequentially in random order. This is a random “local” type of move.

• With probability 0.2 the algorithm proposes the best, non-tabu in the 1-bit flip neighborhood, model (as in TS). The reverse of this move is then added to the tabu list, making such a move forbidden for the next $s$ steps, and therefore possible cycling is avoided.

• With probability 0.1 the algorithm moves to the current best model, an intensification type of move in TS.
Part A (SATS algorithm) (continued)

- With probability 0.3 the algorithm proposes from the $p$, in total, 1-bit flip moves, only the ones that satisfy a **desirability criterion**. We calculate $r_j = \text{cor}(Y, X_j)$ and then the **desirability measure** $d_j = \max r_j / |r_j|$. We transform the desirability measure to probabilities

$$p_j^{\text{in}} = p_{\text{min}} + (p_{\text{max}} - p_{\text{min}}) e^{-c(d_j - 1)},$$

where $(p_{\text{min}}, p_{\text{max}}, c)$ are tuning constants to be specified by the user and $p_j^{\text{out}} = 1 - p_j^{\text{in}}$. Here $p_{\text{min}}$ and $p_{\text{max}}$ govern how dogmatic the inclusion and exclusion processes should be, and $c$ controls the rate at which desirability translates into probability of inclusion. Experimentation resulted in the choices $(p_{\text{min}}, p_{\text{max}}, c) = (0.1, 0.9, 0.1)$.

- If the algorithm gets stuck in the same place for 50 consecutive iterations, a random restart is implemented, as in the **diversification step** of TS.

- In the above algorithm the criterion function is set equal to the **posterior model probability**. The initial temperature is set to $T_0 = 1.0$, the final to $T_f = 0.1$, the reciprocal cooling schedule is used and the tabu list size is set equal to 7 (see Fouskakis and Draper 2008, JASA).
Part B (GA algorithm)

- After running SATS, for a specified number of iterations, the 20 best solutions are returned; those solutions together with 30 randomly chosen configurations form an initial population from where GA starts. In GA, **elitist strategies** are added and therefore good configurations are not getting lost, together with the **highly uniform crossover operator**.

- Starting GA from a set of good configurations, resulting from SATS, will result to an in-depth exploration of this set, with the highly uniform crossover operator to combine efficiently good elements of those configurations. At the end of each iteration of GA the 20 best solutions are returned and a new population is formed by the addition 30 new, randomly chosen, configurations.
7 Experimental Results - Normal Linear Model

\[ Y | X_\gamma, \beta_0, \beta_\gamma, \sigma^2, \gamma \sim N(1_n \beta_0 + X_\gamma \beta_\gamma, \sigma^2 I_n) \]

\[ f(\beta_\gamma | \gamma, \sigma^2) = N\left[0, n\sigma^2 \left(X_\gamma^T X_\gamma\right)^{-1}\right] \]

\[ f(\beta_0, \sigma^2 | \gamma) \propto \sigma^{-2} \]

where \( \beta_0 \) is an intercept that is common to all models, \( \beta_\gamma \) is the \( d_\gamma \)-dimensional vector of nonzero regression coefficients included in the model specified by \( \gamma \), \( \sigma^2 \) is the error variance of any model and \( X_\gamma \) is the data matrix corresponding to model \( \gamma \). We also use the beta-binomial prior on the model space with \( \alpha = \beta = 1 \). The marginal likelihood is given by

\[ f(y | \gamma) \propto \frac{(1 + n)^{\frac{n-1-d_\gamma}{2}}}{\left(1 + n(1 - R_\gamma^2)\right)^{\frac{n-1}{2}}} \]  \hspace{1cm} (19)

where \( R_\gamma^2 \) is the ordinary coefficient of determination of regression model \( \gamma \).
7.1 Simulated Example

We illustrate the proposed method by considering the simulated data-set Nott & Kohn (2005, Biometrika). This data-set consists of \( n = 50 \) observations and \( p = 15 \) covariates. The first 10 covariates are generated from a standardized Normal distribution while

\[
X_{ij} \sim N\left( 0.3X_{i1} + 0.5X_{i2} + 0.7X_{i3} + 0.9X_{i4} + 1.1X_{i5}, 1 \right) \text{ for } j = 11, \ldots, 15, \ i = 1, \ldots, 50
\]

and the response from

\[
Y_i \sim N\left( 4 + 2X_{i1} - X_{i5} + 1.5X_{i7} + X_{i11} + 0.5X_{i13}, 2.5^2 \right), \text{ for } i = 1, \ldots, 50.
\]  \hspace{1cm} (20)

SATS algorithm was run initially for 70 iterations, returning the 20 best models. Then the improved GA version with elitist strategies and highly uniform crossover operator was then run for 70 iterations, resulting together with the previous SATS run to a total of 4,740 utility evaluations. The initial population of GA was formed by the resulting 20 best models from SATS together with 30 randomly chosen configurations.
Table 1: Posterior model probabilities for the 10 best models after running GA for 70 iterations. The last two columns present the posterior model probabilities and posterior odds after the full-enumeration search.

<table>
<thead>
<tr>
<th>$m_j$</th>
<th>Model</th>
<th>Visited from SATS</th>
<th>Posterior Model Probabilities (unnormalized - log scale)</th>
<th>Posterior Model Probabilities (full-search)</th>
<th>Posterior Odds ($m_1$ vs. $m_j$) (full-search)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$X_1 + X_7 + X_{11}$</td>
<td>Yes</td>
<td>4.232876</td>
<td>0.1475</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>$X_1 + X_5 + X_7 + X_{11}$</td>
<td>Yes</td>
<td>3.669204</td>
<td>0.0839</td>
<td>1.7571</td>
</tr>
<tr>
<td>3</td>
<td>$X_1 + X_7$</td>
<td>Yes</td>
<td>3.485628</td>
<td>0.0699</td>
<td>2.1112</td>
</tr>
<tr>
<td>4</td>
<td>$X_1 + X_5 + X_6 + X_7 + X_{11}$</td>
<td>Yes</td>
<td>2.843955</td>
<td>0.0368</td>
<td>4.0105</td>
</tr>
<tr>
<td>5</td>
<td>$X_1 + X_6 + X_7 + X_{11}$</td>
<td>Yes</td>
<td>2.482426</td>
<td>0.0256</td>
<td>5.7572</td>
</tr>
<tr>
<td>6</td>
<td>$X_1 + X_7 + X_{10} + X_{11}$</td>
<td>No</td>
<td>2.152312</td>
<td>0.0184</td>
<td>7.6441</td>
</tr>
<tr>
<td>7</td>
<td>$X_1 + X_4 + X_7$</td>
<td>Yes</td>
<td>2.108140</td>
<td>0.0176</td>
<td>8.3707</td>
</tr>
<tr>
<td>8</td>
<td>$X_1 + X_7 + X_{14}$</td>
<td>No</td>
<td>1.920472</td>
<td>0.0146</td>
<td>10.0987</td>
</tr>
<tr>
<td>9</td>
<td>$X_1 + X_7 + X_{15}$</td>
<td>No</td>
<td>1.855645</td>
<td>0.0137</td>
<td>10.7750</td>
</tr>
<tr>
<td>10</td>
<td>$X_1 + X_7 + X_{11} + X_{12}$</td>
<td>No</td>
<td>1.805890</td>
<td>0.0130</td>
<td>11.3247</td>
</tr>
</tbody>
</table>
7.2 Air-Pollutant Example

We use the ozone data (Breinman & Friedman 1985, JASA) to implement our approaches. The data we used were slightly changed based on some initial preliminary exploratory analysis. As a response we use a standardized version of the logarithm of the ozone concentration variable of the original data set. The standardized versions of nine (9) main effects, 9 quadratic terms, 2 cubic terms, and 36 two-way interactions (a total of 56 variables) where used as possible covariates. The main effects are:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>Day of Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_2$</td>
<td>Wind speed (mph) at LAX</td>
</tr>
<tr>
<td>$X_3$</td>
<td>500 mb pressure height (m) at VAFB</td>
</tr>
<tr>
<td>$X_4$</td>
<td>Humidity (%) at LAX</td>
</tr>
<tr>
<td>$X_5$</td>
<td>Temperature (°F) at Sandburg</td>
</tr>
<tr>
<td>$X_6$</td>
<td>Inversion base height (feet) at LAX</td>
</tr>
<tr>
<td>$X_7$</td>
<td>Pressure gradient (mm Hg) from LAX to Daggett</td>
</tr>
<tr>
<td>$X_8$</td>
<td>Inversion base temperature (°F) at LAX</td>
</tr>
<tr>
<td>$X_9$</td>
<td>Visibility (miles) at LAX</td>
</tr>
</tbody>
</table>
Table 2: The 5 best models after running our two step method for 30+70 iterations (5,476 utility evaluations)

Common variables among the 5 best models: $X_1 + X_2 + X_8 + X_9 + X_{10} + X_{15} + X_{16} + X_{18} + X_{43}$

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Additional Variables</th>
<th>Posterior model Probabilities (unnormalized - log scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>225.7122</td>
</tr>
<tr>
<td>2</td>
<td>$+X_{12}$</td>
<td>224.1090</td>
</tr>
<tr>
<td>3</td>
<td>$+X_7$ $+X_{13}$</td>
<td>223.6418</td>
</tr>
<tr>
<td>4</td>
<td>$+X_{13}$</td>
<td>223.6192</td>
</tr>
<tr>
<td>5</td>
<td>$+X_{12}$ $+X_{20}$</td>
<td>223.4002</td>
</tr>
</tbody>
</table>

For comparison reasons the $MC^3$ algorithm was also run for 100 iterations. In every iteration the $MC^3$ algorithm proposes $p$, in total, 1-bit flip moves (one for each variable) and therefore the final utility evaluations that the algorithm made was 5,600. The models that were ranked in the 2nd, 3rd and 4th place from our two step method were absent from the final top five models of $MC^3$. 
Comparison of the predictive performance

• The posterior distribution of the **deviance statistic**

\[
D(\beta, \gamma) = -2 \sum_{i=1}^{n} \log f(y_i | \beta, \gamma)
\]  

(e.g., Spiegelhalter *et al.* 2002, JRSS B) is often used as a measure of model fit.

Table 3: Comparison of measures of fit and dimensionality between the best model from our approach and the full model; percentage difference is in relation to the full model.

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Full Model</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Deviance</td>
<td>363.2</td>
<td>296.4</td>
<td>+22.5</td>
</tr>
<tr>
<td>Median Deviance</td>
<td>372.9</td>
<td>325.7</td>
<td>+14.5</td>
</tr>
<tr>
<td>Dimension</td>
<td>9</td>
<td>56</td>
<td>-84.0</td>
</tr>
</tbody>
</table>

The minimum and median values of the posterior distribution of the deviance statistic for the best model from our approach were lower by a relatively modest **22.5%** and **14.5%** respectively compared to the corresponding values of the full model, but the dimension of the best model was almost **84%** lower than that for the full model.
Comparison of the predictive performance (continued)

- The **cross-validation log score**, (Geisser and Eddy, 1979, JASA) can also be used as a measure of model fit. It is based on **leave-one-out predictive distributions** $f(y_i|y_{\setminus i})$ and is given by

$$LS_{CV}(\gamma|y) = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i|y_{\setminus i}, \gamma),$$  \hspace{1cm} (22)

where $y_{\setminus i}$ is the vector of data $y$ without observation $i$ (larger values of $LS_{CV}$ indicate greater predictive accuracy). This measure can be estimated directly from a single MCMC run using the formula

$$\hat{LS}_{CV}(\gamma|y) = -\frac{1}{n} \sum_{i=1}^{n} \log \overline{f^{-1}(y_i|\beta, \gamma)},$$  \hspace{1cm} (23)

where $\overline{f^{-1}(y_i|\beta, \gamma)}$ is the posterior mean of the inverse of the predictive density for observation $i$. The $\hat{LS}_{CV}$, for 1,000 MCMC iterations was calculated, for the best and the full models. The values $-0.585$ and $-0.617$ were obtained respectively; **the best model from our approach has an 5.4% larger $\hat{LS}_{CV}$ than the full model.**
8 Experimental Results - Logistic Regression Model

For any model $\gamma \in \mathcal{M}$, of dimension $d_\gamma$ and parameters $\theta_\gamma = (\beta_0, \beta_\gamma)$, the model formulation can be summarized as

$$(Y_i \mid \gamma) \overset{\text{indep}}{\sim} \text{Bernoulli}[p_i(\gamma)],$$

$$\eta_i(\gamma) = \log \left[ \frac{p_i(\gamma)}{1 - p_i(\gamma)} \right] = \beta_0 + \sum_{j=1}^{p} \beta_j \gamma_j X_{ij}$$

$$f(\theta_\gamma \mid \gamma) = N \left[ 0, 4n \left( \tilde{X}_{\gamma}^T \tilde{X}_{\gamma} \right)^{-1} \right]$$

where $\tilde{X}_{\gamma} = (X_{ih}, i = 1, \ldots, n; h = 0, \ldots, p)$ is the design matrix corresponding to model $\gamma$, with $X_{i0} = 1$ for all $i = 1, \ldots, n$. Then the posterior probability of a model $\gamma$ cannot be derived in closed form, and therefore Laplace approximation is used.
\begin{align*}
-2 \log f(\gamma|y) &= -2 \log f(y|\tilde{\theta}_\gamma, \gamma) + [\phi(\gamma) - 2 \log f(\gamma)] + O(n^{-1}), \\
\text{where} \\
\phi(\gamma) &= \frac{1}{4n} \tilde{\theta}_\gamma \tilde{X}_\gamma \tilde{X}_\gamma \tilde{\theta}_\gamma + d\gamma \log(4n) + \log \frac{|\Psi_{\gamma}^{-1}|}{|\tilde{X}_\gamma \tilde{X}_\gamma|} \\
\text{and} \\
\Psi_{\gamma} &= \left[ -\frac{\partial^2 \log f(y|\theta_\gamma, \gamma)}{\partial \theta^2_\gamma} \bigg|_{\theta_\gamma=\tilde{\theta}_\gamma} - \frac{\partial^2 \log f(\theta_\gamma|\gamma)}{\partial \theta^2_\gamma} \bigg|_{\theta_\gamma=\tilde{\theta}_\gamma} \right]^{-1} \\
&= \left( \tilde{X}_\gamma^T \text{diag} \left\{ \exp \left( \tilde{X}_\gamma, i \tilde{\theta}_\gamma \right) \left[ 1 + \exp \left( \tilde{X}_\gamma, i \tilde{\theta}_\gamma \right) \right]^2 + \frac{1}{4n} \right\} \tilde{X}_\gamma \right)^{-1}
\end{align*}

where \( \tilde{X}_{\gamma,i} \) is row \( i \) of the matrix \( \tilde{X}_\gamma \) for \( i = 1, \ldots, n \) and \( \tilde{\theta}_\gamma \) is the posterior mode of \( f(\theta_\gamma|y, \gamma) \).

As before the \textbf{beta-binomial hierarchical prior} is used on the model space with \( \alpha = \beta = 1 \) is considered.
8.1 Quality of Health Care Example

Data from a major US study of quality of hospital care for $n = 2532$ elderly patients suffering from pneumonia were used. The response variable $Y_i$ is equal to 1 if patient $i$ dies within 30 days of admission and 0 otherwise and an initial list of $p = 83$ pneumonia predictors was available. A logistic regression model was used for a construction of a sickness scale. More details on the data-set can be found in Fouskakis and Draper (2008), JASA.

SATS algorithm was run initially for 50 iterations, returning the 20 best models. The improved GA version with elitist strategies and highly uniform crossover operators was then run for 100 iterations. The total number of utility evaluations that the combined algorithm made was 7,844.
Table 4: The 5 best models after running our two step method for 50+100 iterations

Common variables among the 5 best models: $X_1 + X_2 + X_3 + X_4 + X_5 + X_{12} + X_{70} + X_{73}$

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Additional Variables</th>
<th>Posterior model Probabilities (unnormalized - log scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$+X_8$  $+X_{37}$</td>
<td>-861.8985</td>
</tr>
<tr>
<td>2</td>
<td>$+X_8 + X_{15} + X_{37}$</td>
<td>-861.9752</td>
</tr>
<tr>
<td>3</td>
<td>$+X_{37}$</td>
<td>-861.9863</td>
</tr>
<tr>
<td>4</td>
<td>$+X_{15} + X_{37}$</td>
<td>-862.3158</td>
</tr>
<tr>
<td>5</td>
<td>$+X_8$  $+X_{49}$</td>
<td>-864.0107</td>
</tr>
</tbody>
</table>

For comparison reasons the $MC^3$ algorithm was also run, using Laplace approximation, for 95 iterations. As before, in every iteration the $MC^3$ algorithm proposes $p$ in total 1-bit flip moves (one for each variable) and therefore the final utility evaluations that the algorithm made was 7,885. None of the 5 best models from Table 4 were present in the top five models of $MC^3$; the five best models of $MC^3$ were inferior to those on Table 4.
Comparison of the predictive performance

Table 5: Comparison of measures of fit and dimensionality between the best model from our approach and the full model; percentage difference is in relation to the full model.

<table>
<thead>
<tr>
<th></th>
<th>Best Model</th>
<th>Full Model</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Deviance</td>
<td>1579.3</td>
<td>1512.6</td>
<td>+4.2</td>
</tr>
<tr>
<td>Median Deviance</td>
<td>1588.5</td>
<td>1552.3</td>
<td>+2.3</td>
</tr>
<tr>
<td>Dimension</td>
<td>10</td>
<td>83</td>
<td>−88.0</td>
</tr>
</tbody>
</table>

The minimum and median values of the posterior distribution of the deviance statistic for the best model from our approach were lower by a relatively modest 4.2% and 2.3% respectively compared to the corresponding values of the full model, but the dimension of the best model was 88% lower than that for the full model.

The $\hat{L}_{SCV}$ was also calculated, for the best model from our approach and for the full model, resulting to values $−0.316$ and $−0.325$ respectively; the best model from our approach had a 2.8% larger $\hat{L}_{SCV}$ than the full model.
9 Discussion

- A powerful heuristic model search tool was proposed, combining aspects from simulated annealing, genetic algorithm and tabu search.

- The proposed algorithm is a two-step method, combining the information-exchange nature of genetic algorithm with the strong local optimization nature of simulated annealing and the non-cycling nature of tabu search.

- A modified version of simulated annealing was initially performed, using multiple neighborhood structures and features from tabu search and then on the set of best resulting models genetic algorithm with elitist strategies and highly uniform crossover operators was implemented.

- Genetic algorithm proposes moves away from the configurations currently under examination using a neighborhood structure that is completely different from the approach used in simulated annealing or tabu search.

- The proposed combined algorithm can identify optimal or sub-optimal solutions in a limiting amount of CPU time and achieves improved performance when compared with popular MCMC algorithms, such as the MCMC model composition.