

Population-Based Reversible-Jump MCMC for Bayesian Variable Selection and Evaluation Under Cost-Limit Restrictions

Ioannis Ntzoufras,

Department of Statistics, Athens University of Economics and Business, Athens, Greece;
e-mail: ntzoufras@aueb.gr.

Joint work with:

Dimitris Fouskakis & David Draper

Department of Mathematics,
National Technical University of Athens,
Athens, Greece;
e-mail: fouskakis@math.ntua.gr.

Department of Applied Mathematics and
Statistics, University of California,
Santa Cruz, USA;
e-mail: draper@ams.ucsc.edu

Presentation is available at: <http://stat-athens.aueb.gr/~jbn/current.htm>

Synopsis

1. Motivation - The Data.
2. Model Specification.
3. RJMCMC for variable selection.
4. Cost-Restricted Benefit Analysis.
5. Discussion.

1 Motivation

Health care quality measurements

Indirect method: input-output approach.

- Construct a model on hospital outcomes (e.g., mortality within 30 days of admission) *after adjusting for differences in inputs* (sickness at admission).
- Compare observed and expected outcomes to infer for the health care quality.
- Data collection costs are available for each variable (measured in minutes or monetary units).
- We wish to incorporate cost in our analysis in order to reduce data collection costs but also have a well-fitted model.

Available data

- **Data** come from a major U.S. study constructed by RAND Corporation, with $n = 2,532$ pneumonia patients (Keeler, *et al.*, 1990).
- **Response variable:** mortality within 30 days of admission
- **Covariates:** $p = 83$ sickness indicators
- Construct a **sickness scale** using a logistic regression model.
- **Benefit - Only Analysis** (no costs): Classical variable selection techniques to find an “optimal” subset of 10-20 indicators. The initial list of $p = 83$ sickness indicators was reduced to 14 “significant” predictors (Keeler, *et al.*, 1990).

The 14-Variable Rand Pneumonia Scale

The RAND admission sickness scale for pneumonia ($p = 14$ variables), with the marginal data collection costs per patient for each variable (in minutes of abstraction time).

Variable	Cost (Minutes)	Variable	Cost (Minutes)
1 Systolic Blood Pressure Score (2-point scale)	0.5	8 Septic Complications (yes, no)	3.0
2 Age	0.5	9 Prior Respiratory Failure (yes, no)	2.0
3 Blood Urea Nitrogen	1.5	10 Recently Hospitalized (yes, no)	2.0
4 APACHE II Coma Score (3-point scale)	2.5	12 Initial Temperature	0.5
5 Shortness of Breath Day 1 (yes, no)	1.0	17 Chest X-ray Congestive Heart Failure Score (3-point scale)	2.5
6 Serum Albumin Score (3-point scale)	1.5	18 Ambulatory Score (3-point scale)	2.5
7 Respiratory Distress (yes, no)	1.0	48 Total APACHE II Score (36-point scale)	10.0

Two different approaches for incorporating cost into the analysis

Two desirable but opposite criteria must be accounted in our analysis:

1. the fit of the model
2. the cost of the model

Thus, we wish to find a model with the lower possible cost but having an “acceptable fit” to the observed data.

So two different cases for handling cost may appear

Case 1: Decrease the cost as much as possible but without losing much from the predictive ability of the model. No overall budgetary restrictions exist.

Case 2: An overall budgetary bound is implemented. We select the “best” model under the restricted model space.

Three methods for solving this problem

- (1) **Bayesian decision theoretic** solution proposed by Draper and Fouskakis (2000) and Fouskakis and Draper (2002, 2008).
They used stochastic optimization methods to find (near-) optimal subsets of predictor variables that maximize an expected utility function which trades off data collection cost against predictive accuracy [case 1].
- (2) **Model specification using a cost-adjusted prior.** As an alternative to (1), we propose a prior distribution that accounts for the cost of each variable and results in a set of posterior model probabilities. This approach leads to a **generalized cost-adjusted version of the Bayesian Information Criterion** (Fouskakis, Ntzoufras and Draper, 2009a) [case 1].
- (3) **Cost-restriction benefit analysis.** The model search is conducted only among models whose cost does not exceed a **budgetary restriction** (Fouskakis, Ntzoufras and Draper, 2009b), by the usage of a **population-based trans-dimensional RJMCMC method** [case 2].

Cost–restriction benefit analysis

- Select between models whose cost does not exceed a **budgetary restriction** (Fouskakis, Ntzoufras and Draper, 2009b)
- The new truncated model space is more complicated with areas of local maximum (i.e. neighborhoods of “good” models).
- A sophisticated model search algorithm must be adopted.
- We use a **population–based trans–dimensional RJMCMC method**

2 Model Specification

- Logistic regression model with $Y_i = 1$ if patient i dies after 30 days of admission.
- X_{ij} : j sickness predictor variable for the i patient.
- $m \rightarrow \boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^T$.
- γ_j : Binary indicators of the inclusion of the variable X_j in the model.
- Model space $\mathcal{M} = \{0, 1\}^p$; p = total number of variables considered.

Hence the model formulation can be summarized as

$$\begin{aligned}
 (Y_i | \boldsymbol{\gamma}) &\stackrel{\text{indep}}{\sim} \text{Bernoulli}(p_i(\boldsymbol{\gamma})), \\
 \eta_i(\boldsymbol{\gamma}) = \log \left(\frac{p_i(\boldsymbol{\gamma})}{1 - p_i(\boldsymbol{\gamma})} \right) &= \sum_{j=0}^p \beta_j \gamma_j X_{ij}, \\
 \boldsymbol{\eta}(\boldsymbol{\gamma}) &= \mathbf{X} \text{diag}(\boldsymbol{\gamma}) \boldsymbol{\beta} = \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}.
 \end{aligned}$$

3 Reversible Jump MCMC

- Introduced by Green (1995, *Bka*).
- Generalizes simple Metropolis-Hastings algorithms for ‘jumping’ between models of different parameter dimension.
- Estimates posterior model probabilities.
- Can be used for variable selection.
- Extremely flexible and fashionable.
- Complicated in some cases.

The general RJMCMC

- Propose a new model m' with probability $j(m, m')$.
- Generate \mathbf{u} from a specified proposal density $q(\mathbf{u}|\boldsymbol{\beta}_m, m, m')$.
- Set $(\boldsymbol{\beta}'_{m'}, \mathbf{u}') = h_{m, m'}(\boldsymbol{\beta}_m, \mathbf{u})$, where $h_{m, m'}$ is a specified invertible function for which $h_{m', m} = h_{m, m'}^{-1}$. This equation implies that $d_m + d(\mathbf{u}) = d_{m'} + d(\mathbf{u}')$, where d_m and $d(\mathbf{u})$ are the dimensions for the parameters of model m and vector \mathbf{u} , respectively.
- Accept the proposed move to model m' with probability

$$\alpha = \min \left(1, \frac{f(\mathbf{y}|\boldsymbol{\beta}'_{m'}, m')f(\boldsymbol{\beta}'_{m'}|m')f(m')j(m', m)q(\mathbf{u}'|\boldsymbol{\beta}'_{m'}, m', m)}{f(\mathbf{y}|\boldsymbol{\beta}_m, m)f(\boldsymbol{\beta}_m|m)f(m)j(m, m')q(\mathbf{u}|\boldsymbol{\beta}_m, m, m')} \left| \frac{\partial h(\boldsymbol{\beta}_m, \mathbf{u})}{\partial(\boldsymbol{\beta}_m, \mathbf{u})} \right| \right). \quad (1)$$

RJMCMC for Variable Selection

In variable selection m can be substituted by γ .

- Select randomly a variable j and propose it to change.
- If $\gamma_j = 0 \rightarrow \gamma'_j = 1$ (include a new covariate X_j in the model)
 - Propose $\beta'_j = \beta_j \sim q(\beta_j | \boldsymbol{\beta}_\gamma)$
 - Set $\beta'_k = \beta_k$ for all k with $\gamma_k = 1$ ($k \neq j$).
 - Accept the proposed move with probability

$$\alpha = \min(1, O_j) \text{ with } O_j = \frac{f(\mathbf{y} | \boldsymbol{\beta}'_{\gamma'}, \gamma') f(\boldsymbol{\beta}'_{\gamma'} | \gamma') f(\gamma')}{f(\mathbf{y} | \boldsymbol{\beta}_\gamma, \gamma) f(\boldsymbol{\beta}_\gamma | \gamma) f(\gamma) q(\beta_j | \boldsymbol{\beta}_\gamma)}.$$

- If $\gamma_j = 1 \rightarrow \gamma'_j = 0$ (exclude a covariate X_j from the model)
 - Set $\beta'_k = \beta_k$ for all k with $\gamma'_k = 1$ ($k \neq j$).
 - Accept the proposed move with probability

$$\alpha = \min(1, O_j^{-1}).$$

RJMCMC for Variable Selection

- $u = \beta_j$.
- $h(\beta_{\gamma'}) = (\beta_{\gamma}, \beta_j)$ (identity function).
- The Jacobian is equal to one.
- Proposal for each β_j can be a $N(\bar{\mu}_j, \bar{\sigma}_{\beta_j}^2)$ with the mean and variance taken from the posterior of the full model.
- More complicated proposals can be used.
- More complicated model moves can be also used.

4 Cost Restriction - Benefit Analysis

- Implement a **Cost-restriction benefit analysis**, in which the practical relevance of the selected variable subsets is ensured by enforcing an overall limit on the total data collection cost of each subset: the search is conducted only among models whose cost does not exceed this budgetary restriction C .
- Therefore, we should a-priori exclude models γ with total cost larger than C , resulting to a significantly reduced model space,

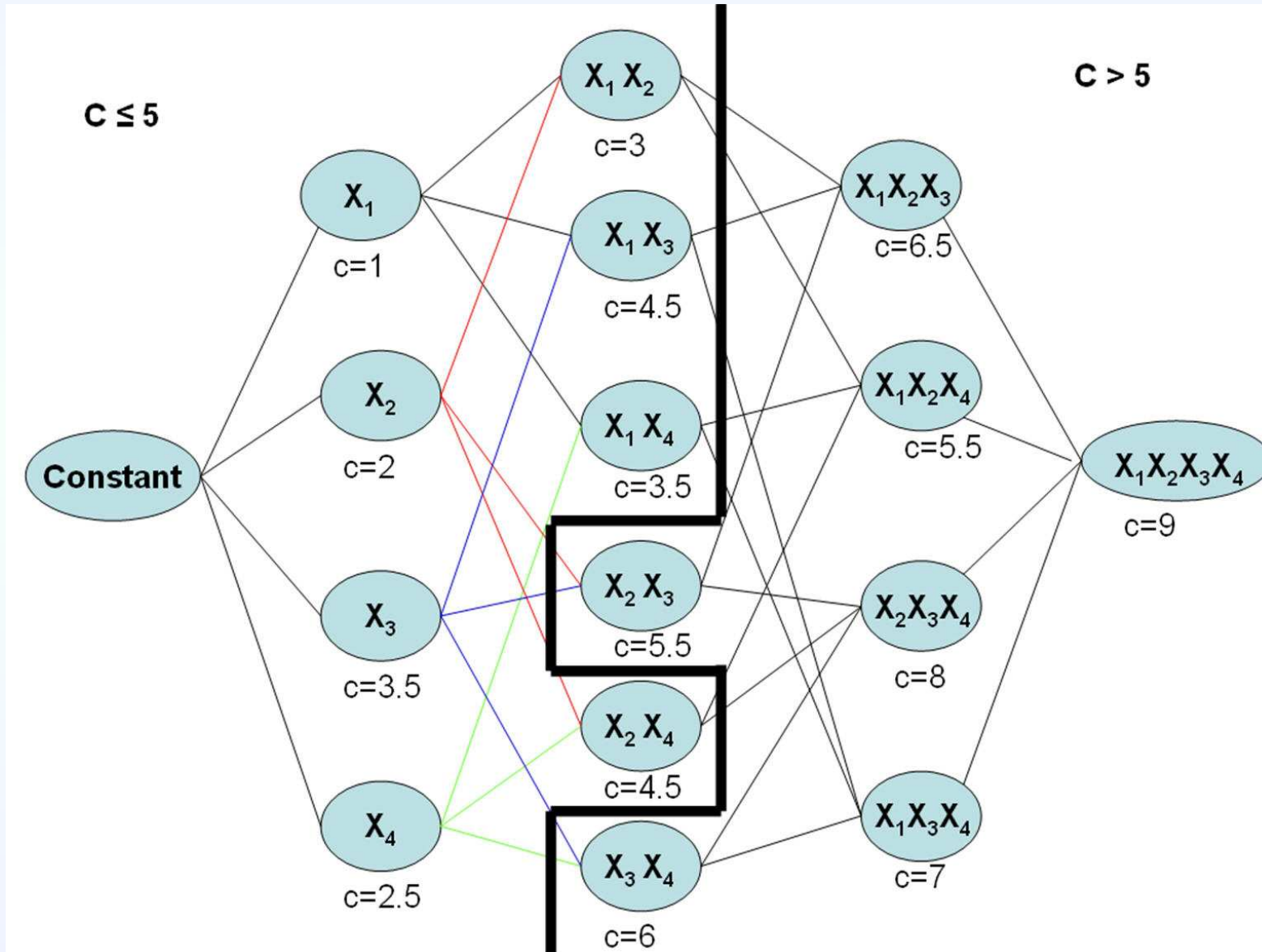
$$\mathcal{M} = \left\{ \gamma \in \{0, 1\}^p : \sum_{j=1}^p c_j \gamma_j \leq C \right\}.$$

- **AIM:** Estimate posterior model probabilities in the cost restricted model space.

PROBLEM

- Due to the cost limit, model space areas of local maximum exist.
- RJMCMC and other Gibbs based samplers for variable selection, move to local model neighborhoods usually by adding or deleting one variable at a time.
- Thus, we need to construct more advanced proposed jumps possibly between models of the same cost in order to avoid getting trapped into local maxima.

Example: Variables X_1 , X_2 , X_3 and X_4 with costs 1, 2, 3.5, 2.5 and total cost limit $C = 5$.



SOLUTION

Intelligent trans-dimension MCMC methods that allow to move across areas of local maximum even if these are distinct.

Proposed Algorithm

We have developed a Population Based Trans-Dimensional Reversible-Jump Markov Chain Monte Carlo algorithm (**Population RJMCMC**), combining ideas from the **population-based MCMC** (Jasra, Stephens and Holmes, 2007) and **Simulated Tempering** (Geyer and Thompson, 1995) algorithms.

Population Based MCMC

- Use k parallel chains (usually around 10) with the posterior raised in a power called (inverse) temperature.
- For temperature $t_1 = 1$ we obtain the actual posterior.
- $0 < t_k < 1$ makes the posterior flatter \Rightarrow move easier between areas of local maxima.
- $t_k > 1$ makes the posterior steeper \Rightarrow move towards local maxima.
- **Problem 1:** How many chains?
- **Problem 2:** Large number of chains is computationally demanding.
- **Problem 3:** What temperatures shall we select?

4.1 The proposed population based algorithm

Population RJMCMC (1)

- Use **3 chains**: The actual one, plus two **auxiliary** ones.
 - In the auxiliary chains the posterior distributions are raised in a power t_k (**inverse temperature**), $k = 1, 2$.
 - **1st auxiliary chain**: $t_1 > 1 \rightarrow$ increasing differences between the posterior probabilities (makes the distribution steeper \Rightarrow move closer to locally best models).
 - **2nd auxiliary chain**: $0 < t_2 < 1 \rightarrow$ reducing differences between the posterior probabilities (makes the distribution flatter \Rightarrow move easily across different models).

- Inverse temperatures t_k change stochastically.
- By this way the extensive number of chains is avoided (usually from 5-10 in population based samplers).
- The **incorporation of stochastic temperatures** can be done using pseudo priors $g_k(t_k)$.

Population RJMCMC (2)

- The posterior distribution is expanded to

$$\begin{aligned}
 & f(\boldsymbol{\beta}_\gamma, \gamma, \boldsymbol{\beta}_{\gamma,(1)}, \gamma_{(1)}, \boldsymbol{\beta}_{\gamma,(2)}, \gamma_{(2)}, t_1, t_2 | \mathbf{y}) \\
 & \propto f(\mathbf{y} | \boldsymbol{\beta}_\gamma, \gamma) f(\boldsymbol{\beta}_\gamma | \gamma) f(\gamma) \\
 & \quad \times \prod_{k=1}^2 \left\{ f(\mathbf{y} | \boldsymbol{\beta}_{\gamma,(k)}, \gamma_{(k)}) f(\boldsymbol{\beta}_{\gamma,(k)} | \gamma_{(k)}) f(\gamma_{(k)}) \right\}^{t_k} g_k(t_k),
 \end{aligned}$$

where $\gamma_{(k)}$ and $\boldsymbol{\beta}_{\gamma,(k)}$ are the model indicator and parameter vector of chain k .

- Model indicators and parameters can be updated using RJMCMC steps.

Population RJMCMC (2)

- In Gibbs sampling, the inverse temperature t_k is generated by

$$f(t_k | \beta, \gamma, \beta_{\gamma, (k)}, \gamma_{(k)}, t_{\setminus k}, \mathbf{y}) \propto \left\{ f(\mathbf{y} | \beta_{\gamma, (k)}, \gamma_{(k)}) f(\beta_{\gamma, (k)} | \gamma_{(k)}) f(\gamma_{(k)}) \right\}^{t_k} g_k(t_k).$$

PROBLEM: When flat (non informative prior) for inverse temperatures is imposed then the conditional distribution above is an increasing function of t_k .

SOLUTION: The temperatures are only used to expand the space and to make possible jumps between models of different dimension and structure. So $g_k(t_k)$ are not actual priors but pseudo-priors.

- We propose to use directly the marginal posterior distribution of the inverse temperatures t_k $f(t_k | \mathbf{y})$ in the sampling scheme.
- The desired posterior marginal distribution for the inverse temperatures t_k is

given by

$$\begin{aligned}
 f(t_k | \mathbf{y}) &\propto \sum_{\gamma_{(k)} \in \mathcal{M}} \int_{\beta_{\gamma_{(k)}}} \left(f(\mathbf{y} | t_k, \beta_{\gamma_{(k)}}, \gamma_{(k)}) f(\beta_{\gamma_{(k)}} | \gamma_{(k)}) f(\gamma_{(k)}) \right)^{t_k} g_k(t_k) d\beta_{\gamma_{(k)}} \\
 &\propto Z_k(\mathbf{y}, t_k) g_k(t_k),
 \end{aligned}$$

where $Z_k(\mathbf{y}, t_k)$ is the marginal likelihood over all possible models for chain k .

- Since $g_k(t_k)$ are pseudo-priors, we can set

$$g_k(t_k) \propto \frac{h_k(t_k)}{Z_k(\mathbf{y}, t_k)}$$

where $h_k(t_k)$ are convenient and easy to simulate from density functions resulting to

$$f(t_k | \mathbf{y}) = h_k(t_k).$$

- For the selection of $h_k(t_k)$ we propose to use

$$h_1(t_1) = \text{Gamma}(t_1 - 1; a_1, b_1) \text{ and } h_2(t_2) = \text{Beta}(t_2; a_2, b_2).$$

Population RJMCMC (3)

Our algorithm can be summarised as follows:

1. Select initial values for $(\beta_\gamma, \beta_{\gamma,(1)}, \beta_{\gamma,(2)})$ and $(\gamma, \gamma_{(1)}, \gamma_{(2)})$.
2. For $l = 1, \dots, L$ (where L is the number of iterations), repeat:
 - (a) Generate t_1 and t_2 from $f(t_1|\mathbf{y}) = h_1(t_1)$ and $f(t_2|\mathbf{y}) = h_2(t_2)$, respectively.
 - (b) For $k = 0, 1, 2$:
 - i. Sample $\beta_{\gamma,(k)}$ using Gibbs steps.
 - ii. Sample $\gamma_{(k)}$ using RJMCMC steps by proposing to change each component sequentially; thus, for every $j \in \{1, \dots, p\}$ (in a random scan):
 - A. With probability 1 propose $\gamma'_{(k)}$: $\gamma'_{j,(k)} = 1 - \gamma_{j,(k)}$ and $\gamma'_{\ell,(k)} = \gamma_{\ell,(k)}$ for all $\ell \neq j$.
 - B. If $\gamma_{j,(k)} = 1$ then propose $\beta'_{j,(k)}$ from $q_{j,k}(\beta'_{j,(k)})$ and set $\beta'_{\ell,(k)} = \beta_{\ell,(k)}$ for $\ell \neq j$.

C. Accept the proposed move with probability $\alpha = \min\{1, O\}$, where

$$O = \left[\frac{f(\mathbf{y}|\boldsymbol{\beta}'_{\boldsymbol{\gamma},(k)}, \boldsymbol{\gamma}'_{(k)}) f(\boldsymbol{\beta}'_{\boldsymbol{\gamma},(k)}|\boldsymbol{\gamma}'_{(k)}) f(\boldsymbol{\gamma}'_{(k)})}{f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}, \boldsymbol{\gamma}_{(k)}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}|\boldsymbol{\gamma}_{(k)}) f(\boldsymbol{\gamma}_{(k)})} \right]^{t_k} \frac{q_{j,k}(\boldsymbol{\beta}_{j,(k)})^{\gamma_{j,(k)}}}{q_{j,k}(\boldsymbol{\beta}'_{j,(k)})^{1-\gamma_{j,(k)}}}. \quad (2)$$

In the above steps, $\boldsymbol{\beta}_{\boldsymbol{\gamma},(0)}$ and $\boldsymbol{\gamma}_{(0)}$ correspond to the parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ of the original chain, and $t_0 = 1$ is the (inverse) temperature of the original chain.

(c) For $k = 1, 2$:

- i. Propose with probability 1 to swap $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \leftrightarrow (\boldsymbol{\beta}_{(k)}, \boldsymbol{\gamma}_{(k)})$.
- ii. Accept the proposed move with probability $\alpha = \min\{1, O\}$, where

$$O = \left[\frac{f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}, \boldsymbol{\gamma}_{(k)}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}|\boldsymbol{\gamma}'_{(k)}) f(\boldsymbol{\gamma}'_{(k)})}{f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma})} \right]^{1-t_k}. \quad (3)$$

The above sampling scheme can be enriched with additional moves used in population MCMC (such as mutation and crossover).

In our problem: the moves described above were sufficient to achieve good mixing.

Proposal Specification

We use Gaussian independent proposals of the form

$$q_{j,k}(\beta_{j,(k)}) \sim \text{Normal} \left[\bar{\beta}_{j,(k)}, \bar{\sigma}_{j,(k)}^2 \right]. \quad (4)$$

Proposal parameters $\bar{\beta}_{j,(0)}$ and $\bar{\sigma}_{j,(0)}^2$ for the original chain (as in RJMCMC)

- From posterior summaries of the full chain using a pilot run
- MLEs of the full model
- Conditional maximization estimates.
- more sophisticated techniques (Brooks *et al.* , 2003)

For the other two chains we set

$$\bar{\beta}_{j,(k)} = \bar{\beta}_{j,(0)} \quad \text{and} \quad \bar{\sigma}_{j,(k)}^2 = \frac{\bar{\sigma}_{j,(0)}^2}{t_k} \quad \text{for } k = 1, 2. \quad (5)$$

Why?

Let us consider the following approximation to the posterior distribution:

$$f(\boldsymbol{\beta}_{(k)} | \mathbf{y}) \cong \text{Normal} \left\{ \tilde{\boldsymbol{\beta}}_{(k)}, \left[-H_k \left(\tilde{\boldsymbol{\beta}}_{(k)} \right) \right]^{-1} \right\}, \quad (6)$$

- $\Lambda_k(\boldsymbol{\beta}_{(k)}) = \text{constant} + t_k \log f(\mathbf{y} | \boldsymbol{\beta}_{(k)}, \boldsymbol{\gamma}_{(k)}) + t_k \log f(\boldsymbol{\beta}_{(k)} | \boldsymbol{\gamma}_{(k)}) + t_k \log f(\boldsymbol{\gamma}_{(k)})$ is the heated/cooled log-posterior density under t_k .

- $\tilde{\boldsymbol{\beta}}_{(k)}$ is the value maximizing $\Lambda_k(\boldsymbol{\beta}_{(k)})$.

- $H_k(\tilde{\boldsymbol{\beta}}_{(k)})$ is the Hessian of $\Lambda_k(\boldsymbol{\beta}_{(k)})$ evaluated at its maximum.

\Rightarrow The heated/cooled posterior mode $\tilde{\boldsymbol{\beta}}_{(k)}$ is equal to the mode $\tilde{\boldsymbol{\beta}}$ of the actual posterior density (with temperature 1).

$\Rightarrow H_k(\tilde{\boldsymbol{\beta}}_{(k)}) = t_k H(\tilde{\boldsymbol{\beta}}); \Rightarrow \text{Var}(\tilde{\boldsymbol{\beta}}_{(k)}) = t_k^{-1} \text{Var}(\tilde{\boldsymbol{\beta}});$

⇒ thus the proposal variance of the chain with inverse temperature t_k can be defined as the variance of the chain with temperature 1 divided by t_k , as in the second part of (5).

4.2 Prior Distributions

Prior on model parameters

$$\beta_{\gamma} | \gamma \sim \text{Normal} \left(\mathbf{0}, 4n \left(\mathbf{X}_{\gamma}^T \mathbf{X}_{\gamma} \right)^{-1} \right)$$

- Low information prior defined by Ntzoufras, Delaportas and Forster (2003).
- Can be derived using the power prior of Chen *et al.* (2000) and imaginary data supporting the simplest model included in our model space.
- It gives weight to the prior equal to one data-point.
- It is equivalent to the Zellner's g -prior (with $g = 4n$) used for normal regression models.

Prior on model parameters - details

Assume imaginary data $\mathbf{y}_i^* = (y_{i1}^* = 1, y_{i2}^* = 0), i = 1, \dots, n$ with the same design matrix \mathbf{X}_γ as the actually observed data.

Prior generated using the likelihood of these imaginary data,

$$f(\boldsymbol{\beta}_\gamma | \gamma, \mathbf{y}^*) \propto \left\{ \prod_{i=1}^n p_i(\gamma) [1 - p_i(\gamma)] \right\}^{1/(2n)}, \quad (7)$$

where $\mathbf{y}^* = (\mathbf{y}_1^*, \dots, \mathbf{y}_n^*)$.

Based on the power prior of Chen *et al.* (2000, *JSPI*).

Using the above prior, the posterior becomes

$$f(\boldsymbol{\beta}_\gamma | \gamma, \mathbf{y}) \propto \prod_{i=1}^n p_i(\gamma)^{y_i + \frac{1}{2n}} [1 - p_i(\gamma)]^{(1 + \frac{1}{n}) - (y_i + \frac{1}{2n})}; \quad (8)$$

Equivalent to information from $\sum_{i=1}^n (1 + \frac{1}{n}) = (n + 1)$ data points.

⇒ The prior introduces additional information equivalent to adding one data point to the likelihood

⇒ Support *a priori* the simplest model with a weight of one data point.

Using a Laplace approximation we obtain

$$f(\boldsymbol{\beta}_\gamma | \gamma, \mathbf{y}^*) \sim N \left[\hat{\boldsymbol{\beta}}_\gamma, 2n \mathcal{I}(\hat{\boldsymbol{\beta}}_\gamma)^{-1} \right], \quad (9)$$

- $\hat{\boldsymbol{\beta}}_\gamma$ is the MLE if the imaginary data \mathbf{y}_i^* were observed
- $\mathcal{I}(\hat{\boldsymbol{\beta}}_\gamma)$ is the observed information matrix given by

$$\mathcal{I}(\hat{\boldsymbol{\beta}}_\gamma) = \mathbf{X}_\gamma^T \text{diag} \{2 \hat{p}_i^*(\gamma) [1 - \hat{p}_i^*(\gamma)]\} \mathbf{X}_\gamma, \quad (10)$$

- $\hat{p}_i^*(\gamma) = \left[1 + \exp(-\mathbf{X}_i \hat{\boldsymbol{\beta}}_\gamma) \right]^{-1}$ is the fitted success probability for all i under model γ when observing data \mathbf{y}^* .

Under the imaginary data $\mathbf{y}_i^* = (y_{i1}^* = 1, y_{i2}^* = 0), i = 1, \dots, n$

- $\hat{\boldsymbol{\beta}}_\gamma = \mathbf{0}$.
- $\hat{p}_i(\gamma) = 1/2$ for all i ,

$\Rightarrow \mathcal{I}(\hat{\boldsymbol{\beta}}_\gamma) = \frac{1}{2} \left(\mathbf{X}_\gamma^T \mathbf{X}_\gamma \right) \Rightarrow$ leading to the proposed prior.

This approach is also sensible in terms of the parsimony principle.

Posterior model odds (and Bayes factors) penalize the model likelihood for deviations of the actual data from the prior distribution (see Raftery, 1996, equation 12).

Since the above prior can be generated using a set of minimally-weighted imaginary data that fully support the constant model, it will provide sensible minimal *a priori* support for more parsimonious models.

Prior on model space

Uniform prior on cost restricted model space, i.e.

$$f(\gamma) \propto I \left(\gamma \in \mathcal{M} : c(\gamma) = \sum_{j=1}^p \gamma_j c_j \leq C \right),$$

where c_j is the differential cost per observation for variable X_j and C is the budgetary restriction.

4.3 Implementation and Results

Implementation details

- COST LIMIT: $C = 10$ minutes of abstraction time.
- The Procedure:
 1. Run Population RJMCMC for 100K iterations in the full model space, twice, starting each time from a different model.
 2. Eliminate non-important variables (with marginal probabilities < 0.30 in both runs) forming a new reduced model space.
 3. Run population RJMCMC in the reduced space, twice.
- The pseudo-parameters were tuned to achieved acceptance rates around 20% for swapping values between chains of different temperatures, resulting in $h_1(t_1) = \text{Gamma}(t_1 - 1; 2, 4)$ and $h_2(t_2) = \text{Beta}(t_2; 7, 3)$
- **Population vs. simple RJMCMC:** Comparison of results and performance.

Preliminary Results: Marginal Probabilities $f(\gamma_j = 1|\mathbf{y})$

Variables with marginal posterior probabilities $f(\gamma_j = 1|\mathbf{y})$ above 0.30 in at least one run.

Variable Index	Variable Name	Variable Cost	Marginal Posterior Probabilities	
			First Run Analysis	Second Run Analysis
1	Systolic Blood Pressure (SBP) Score	0.50	0.98	0.99
2	Age	0.50	0.97	0.95
3	Blood Urea Nitrogen	1.50	0.99	0.91
4	Apache II Coma Score	2.50	0.55	1.00
5	Shortness of Breath Day 1	1.00	0.92	0.80
6	Serum Albumin	1.50	0.40	0.55
12	Initial Temperature	0.50	0.91	0.93
37	Apache Respiratory Rate Score	1.00	0.72	0.79
46	Admission SBP	0.50	0.45	0.25
49	Respiratory Rate Day 1	0.50	0.35	0.25
51	Confusion Day 1	0.50	0.44	0.01
62	Body System Count	2.50	0.55	0.33
70	Apache pH Score	1.00	0.81	0.73

Reduced Model Space: Posterior Model Probabilities/Odds

Common variables in both analyses: $X_2 + X_4$

Population RJMCMC - 500K iterations

k	m	Common Variables	Additional Variables	1st Run		2nd Run		
				Posterior Prob.	PO_{1k}^*	Posterior Prob.	PO_{1k}^*	
1	m_1	$X_1 + X_{12} + X_{37}$	$+X_3 + X_5$	$+X_{62}$	0.4872	1.00	0.4879	1.00
2	m_2		$+X_5$	$+X_{46} + X_{62} + X_{70}$	0.1202	4.05	0.1052	4.63
3	m_3		$+X_3$	$+X_{62} + X_{70}$	0.0894	5.45	0.0982	4.97
4	m_4		$+X_3 + X_5 + X_6$	$+X_{70}$	0.0344	14.16	0.0498	9.80

Simple RJMCMC - 1500K iterations

k	m	Common Variables	Additional Variables	1st Run		2nd Run			
				Posterior Prob.	PO_{1k}^*	Posterior Prob.	PO_{1k}^*		
1	m_1	X_{62}	$+X_1 + X_3 + X_5 + X_{12} + X_{37}$		0.6159	1.00	0.5912	1.00	
2	m_3		$+X_1 + X_3$	$+X_{12} + X_{37}$	$+X_{70}$	0.1061	5.80	0.1525	3.88
3	m_2		$+X_1$	$+X_5 + X_{12} + X_{37} + X_{46}$	$+X_{70}$	0.0926	6.65	0.1041	5.68
4	m_5		$+X_3 + X_5$	$+X_{46} + X_{49} + X_{70}$		0.0403	15.28	< 0.03	> 19.9

* posterior odds of the best model within each analysis versus the current model k .

All models appearing in the table have total cost 10 min (cost limit).

Reduced Model Space: Monte Carlo Errors

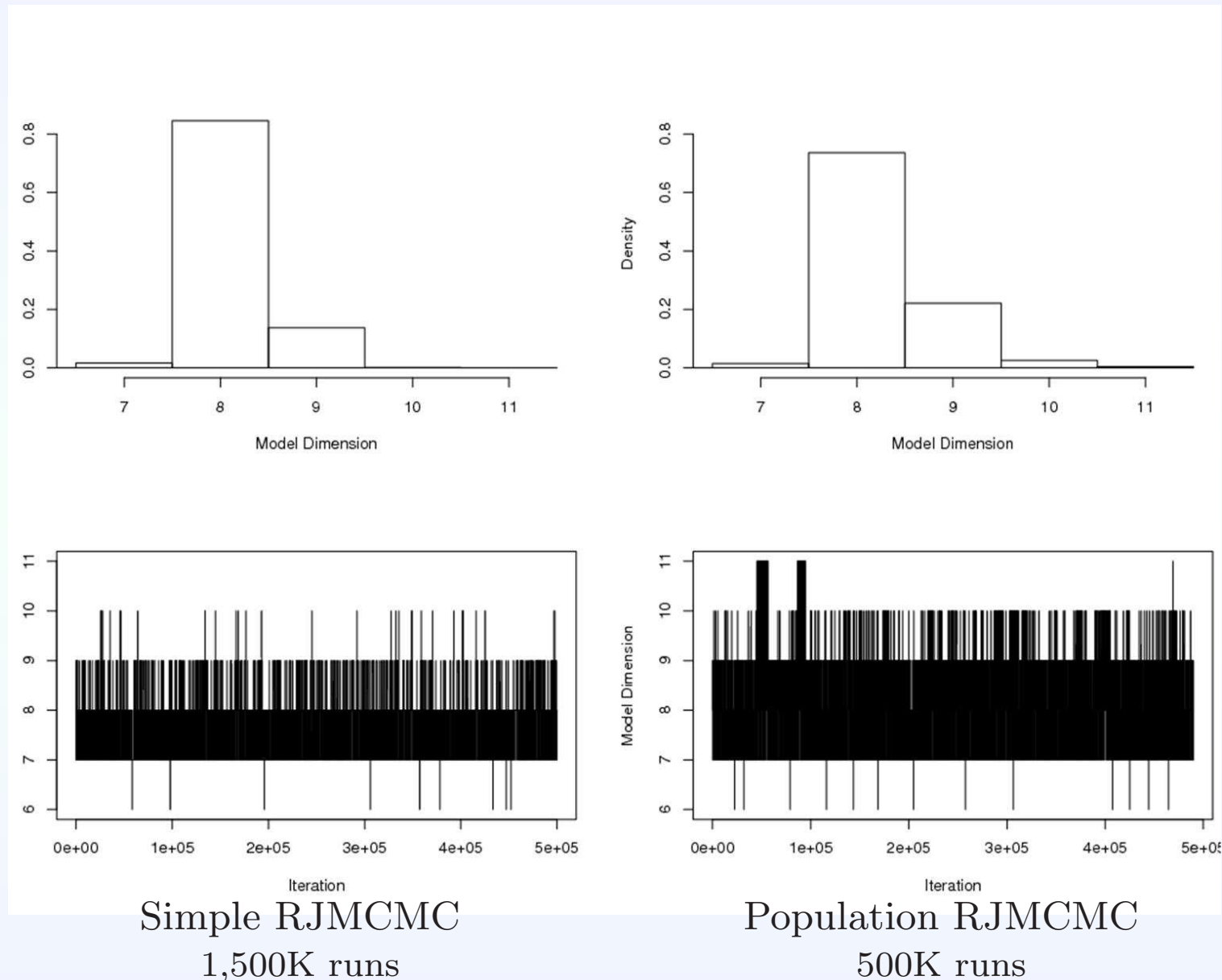
RJMCMC			MCSEs (%)			
Type	Run	Iterations	m_1	m_2	m_3	m_4
<i>P</i>	1	500K	1.2	0.5	0.9	0.7
<i>P</i>	2	500K	1.5	0.4	1.0	0.7
<i>P</i>	1	200K	1.9	0.8	1.1	1.2
<i>P</i>	2	200K	1.6	1.0	1.1	0.9
<i>P</i>	1	100K	2.5	1.2	1.7	1.5
<i>P</i>	2	100K	2.7	0.9	1.6	1.2
<i>S</i>	1	500K	4.2	1.3	3.2	0.0
<i>S</i>	2	500K	4.2	1.7	3.6	0.0
<i>S</i>	1	1,500K	2.9	1.1	2.1	1.0
<i>S</i>	2	1,500K	3.1	0.9	3.1	0.0

<i>P</i> Iterations		Relative Comparisons			
First 1,500K <i>S</i> Run versus <i>P</i>	500K	2.4	2.2	2.3	1.4
	200K	1.5	1.4	1.9	0.8
	100K	1.2	0.9	1.2	0.7
Second 1,500K <i>S</i> Run versus <i>P</i>	500K	2.1	2.3	3.1	0.0
	200K	1.9	0.9	2.8	0.0
	100K	1.2	1.0	1.9	0.0

Comparison of the best models and the RAND model

Model	Minimum Deviance	Total Cost	Dimension
m_1	1610.0	10	8
m_2	1606.7	10	9
m_3	1612.8	10	8
m_4	1608.6	10	9
m_5	1616.5	10	8
RAND	1587.3	31	14
Bayesian Benefit	1553.2	22.5	13

Figure 1: Density and time series plots of model dimension.



Simple RJMCMC

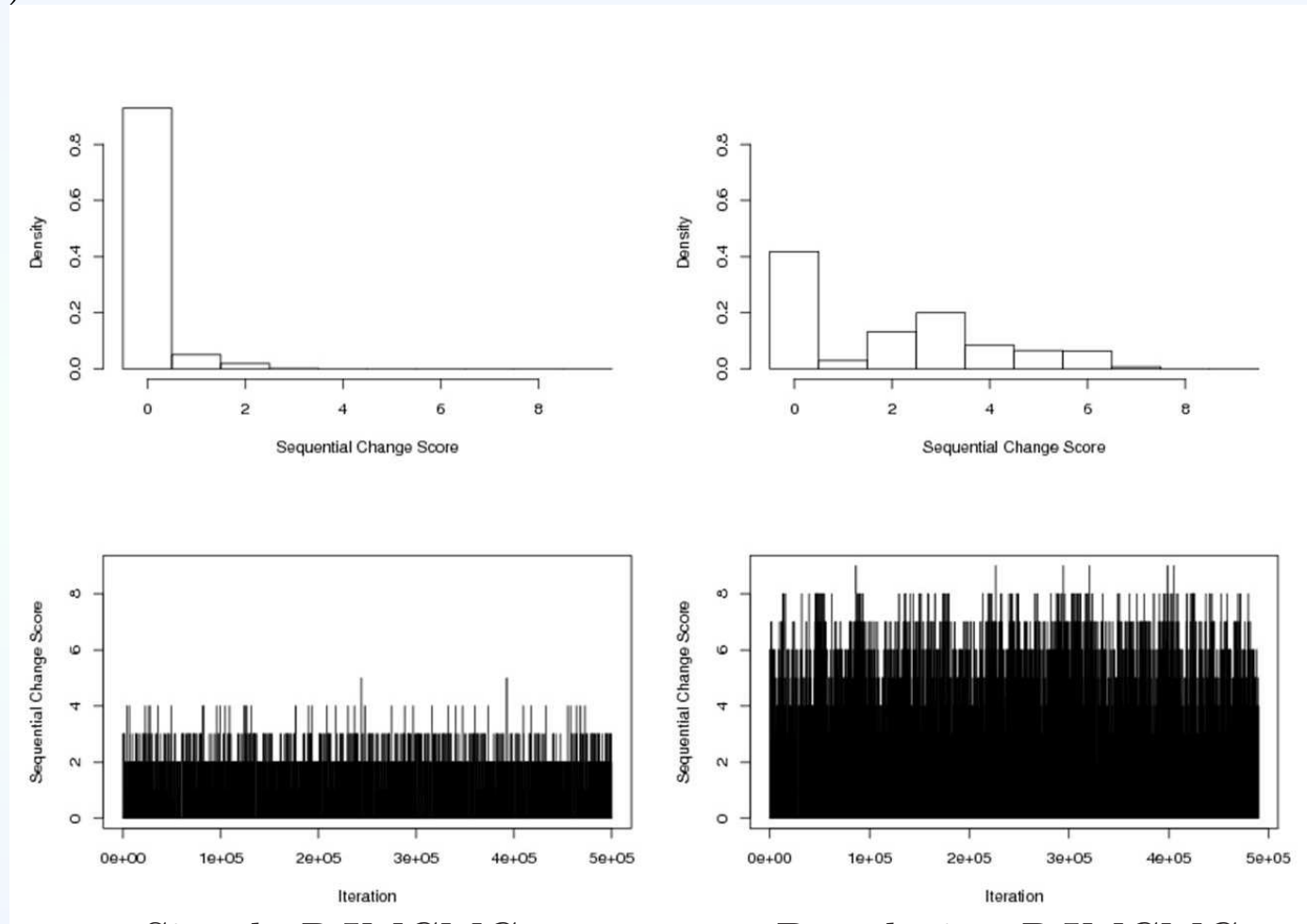
1,500K runs

(thinned by a factor of 3)

Population RJMCMC

500K runs

Figure 2: Density and time series plots of sequential change score (the number of variables in the model at iteration $(t + 1)$ that are different from those in the model at iteration t)



Simple RJMCMC

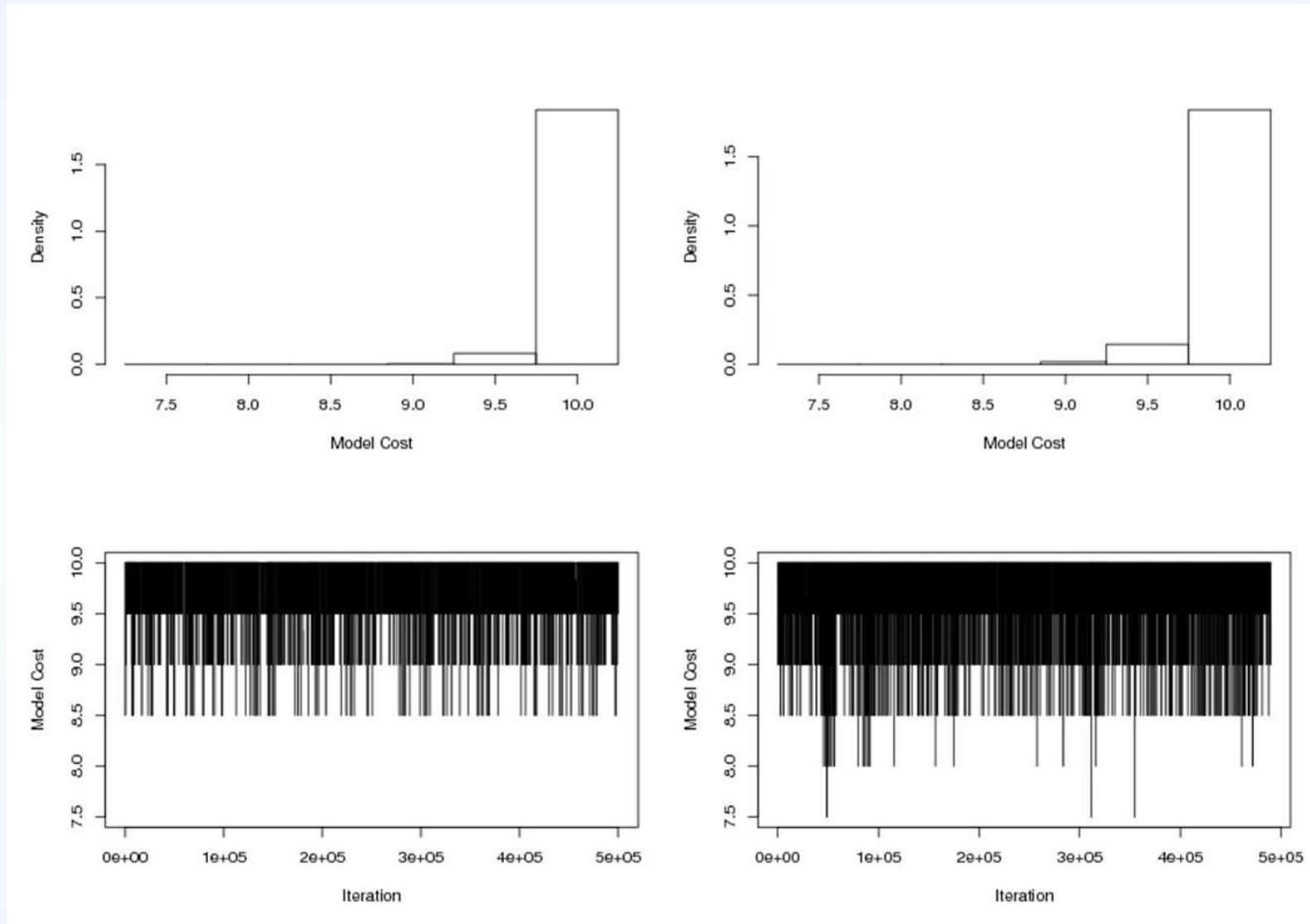
1,500K runs

(thinned by a factor of 3)

Population RJMCMC

500K runs

Figure 3: Density and time series plots of model cost



Simple RJMCMC

1,500K runs

(thinned by a factor of 3)

Population RJMCMC

500K runs

5 Discussion

- **Cost - Restriction - Benefit Analysis:**

A modified Population RJMCMC algorithm is proposed to explore the restricted model space when budgetary constraints are imposed.

The proposed algorithm explores the model space efficiently and converges faster than simple RJMCMC (having lower Monte Carlo errors).

Authors' related work

Fouskakis D, Ntzoufras I, Draper D (2009b). Population Based Reversible Jump MCMC for Bayesian Variable Selection and Evaluation Under Cost Limit Restrictions. *Journal of the Royal Statistical Society C (Applied Statistics)* **58**, 383–403.

Additional References

- Geyer CJ, Thomson EA (1995). Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90**, 909–920.
- Green P (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Jasra A, Stephens DA, Holmes CC (2007). Population-based reversible jump MCMC. *Biometrika*, **94**, 787–807.
- Keeler E, Kahn K, Draper D, Sherwood M, Rubenstein L, Reinisch E, Kosecoff J, Brook R (1990). Changes in sickness at admission following the introduction of the Prospective Payment System. *Journal of the American Medical Association*, **264**, 1962–1968.
- Ntzoufras I, Dellaportas P, Forster JJ (2003). Bayesian variable and link determination for generalized linear models. *Journal of Statistical Planning and Inference*, **111**, 165–180.