Incorporating cost in Bayesian variable selection, with application to cost-effective measurement of quality of health care

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. Cost - Benefit Analysis.
4. Cost - Restriction - 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 form 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, 2008a) [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, 2008b), by the usage of a **population–based trans–dimensional RJMCMC method** [case 2].

Here we present results from methods (2) and (3).

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.
- $m \rightarrow \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$.
- 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

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

3 Bayesian Cost-Benefit Analysis

The aim is to identify well fitted models after taking into account the cost of each variable.

Therefore we need to estimate the posterior model probability

$$f(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{f(\boldsymbol{\gamma}) \int f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}}{\sum_{\boldsymbol{\gamma}' \in \{0,1\}^p} f(\boldsymbol{\gamma}') \int f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}'}, \boldsymbol{\gamma}') f(\boldsymbol{\beta}_{\boldsymbol{\gamma}'}|\boldsymbol{\gamma}') d\boldsymbol{\beta}_{\boldsymbol{\gamma}'}}$$

after introducing a prior on model space $f(\boldsymbol{\gamma})$ depending on the cost.

Why this approach instead of decision theoretic

- **Bayesian Decision Theoretic** (Draper and Fouskakis, 2000; Fouskakis and Draper, 2002, 2008).
  - Demands extensive and detailed specification of the utility function.
  - Stochastic optimization methods must be implemented to identify optimal or near-optimal subsets of "good" covariates.
  - We cannot account for model uncertainty.
- **Proposed approach: Cost Adjusted Prior Model Specification.**
  - Advanced MCMC methods can be used to efficiently search the model space.
  - We can account for model uncertainty via posterior model probabilities and model averaging. A subset of "good" and economical models can be finally accepted for further analysis.
  - A prior distribution is proposed satisfying specific criteria defined a-priori.
  - Resulted posterior model odds can be approximated by a **generalized cost–adjusted BIC**.

3.1 Preliminaries: Posterior model odds and penalty functions

Information criteria (1)

Information criterion for model $\boldsymbol{\gamma}$

$$IC(\boldsymbol{\gamma}) = -2 \log f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) + d_{\boldsymbol{\gamma}} F$$

- $f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})$ is the maximum likelihood.
- $d_{\boldsymbol{\gamma}}$ dimension of the model (number of parameters)
- F penalty for each model parameter used/estimated.
- $d_{\boldsymbol{\gamma}} F$ is the total penalty implemented to the maximum likelihood due to the use of a model with $d_{\boldsymbol{\gamma}}$ parameters.

Model with minimum IC is indicated as the "best".

The above criterion is a penalized likelihood measure.

Information criteria (2)

When comparing two models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)}$ then

$$\begin{aligned} IC_{k\ell} = IC(\boldsymbol{\gamma}^{(k)}) - IC(\boldsymbol{\gamma}^{(\ell)}) &= -2\log \frac{f\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)}\right)}{f\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)}\right)} + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)F \\ &= Deviance_{k\ell} + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)F \end{aligned}$$

We select model $\boldsymbol{\gamma}^{(k)}$ if $IC_{k\ell} < 0$, and model $\boldsymbol{\gamma}^{(\ell)}$ if $IC_{k\ell} > 0$.

Posterior model probabilities and information criteria

The posterior model probability of a model $\boldsymbol{\gamma}$ is given by

$$f(\boldsymbol{\gamma}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})$$

where

- $f(\boldsymbol{y}|\boldsymbol{\gamma})$ is the marginal likelihood of model $\boldsymbol{\gamma}$ given by $\int f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}$
- $f(\boldsymbol{\gamma})$ prior probability of model $\boldsymbol{\gamma}$

It can be rewritten as

$$\begin{array}{ccccc} -2\log f(\boldsymbol{\gamma}|\boldsymbol{y}) & = & -2\log f(\boldsymbol{y}|\boldsymbol{\gamma}) & + & [-2\log f(\boldsymbol{\gamma})] \\ \Updownarrow & & \Updownarrow & & \Updownarrow \\ IC(\boldsymbol{\gamma}) & = & -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) & + & d_{\boldsymbol{\gamma}} F \end{array}$$

Posterior model odds and information criteria

Similarly if we consider the posterior odds of model $\boldsymbol{\gamma}^{(k)}$ versus model $\boldsymbol{\gamma}^{(\ell)}$. Then

$$\begin{aligned} PO_{k\ell} &= \left(\frac{f(\boldsymbol{y}|\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\boldsymbol{\gamma}^{(\ell)})}\right) \times \frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} \\ &= B_{k\ell} \times \mathrm{PrO}_{k\ell}, \end{aligned}$$

- $B_{k\ell}$ is the Bayes factor of model $\boldsymbol{\gamma}^{(k)}$ versus model $\boldsymbol{\gamma}^{(\ell)}$ (ratios of marginal likelihoods).
- $\mathrm{PrO}_{k\ell}$ is the prior odds of model $\boldsymbol{\gamma}^{(k)}$ versus model $\boldsymbol{\gamma}^{(\ell)}$.

It can be rewritten as

$$\begin{array}{ccccc} -2\log PO_{k\ell} = & & -2\log B_{k\ell} & + & 2\log PrO_{\ell k} \\ \Updownarrow & & \Updownarrow & & \Updownarrow \\ IC_{k\ell} & = -2\log \dfrac{f\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)}\right)}{f\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)}\right)} & + & \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)F \end{array}$$

Uniform prior on model space

If the prior model probabilities are defined via a negative function of the model dimension, then the prior model odds

$$\xi_{k\ell} = -2 \log PrO_{k\ell} = -2 \log \frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})}$$

can be also interpreted as the **extra penalty imposed to the Bayes factor** .

If the (usual) uniform prior distribution is used then

$$\xi_{k\ell} = 0 \text{ and } PO_{k\ell} = B_{k\ell} \text{ for all models } \boldsymbol{\gamma}_k, \boldsymbol{\gamma}_\ell \in \mathcal{M}$$

where $\mathcal{M}$ is the set of all models under consideration (model space).

Bayesian benefit-only analysis can be assumed using the uniform prior on model space and hence base our variable selection procedure in Bayes factors.

Prior model odds interpretation

Well-known rough approximation of $\log B_{k\ell}$ (Schwartz, 1978):

$$\begin{aligned} -2\log B_{k\ell} &= BIC_{k\ell} + O(1) \Leftrightarrow \\ -2\log PO_{k\ell} &= BIC_{k\ell} + \xi_{k\ell} + O(1) \\ &= Deviance_{k\ell} + (d_{\boldsymbol{\gamma}^{(k)}} - \boldsymbol{\gamma}^{(\ell)})\log n + \xi_{k\ell} + O(1) \end{aligned} \qquad (1)$$

where $BIC_{k\ell}$ is the Bayesian Information Criterion (e.g., Kass and Wasserman, 1996; Raftery, 1995, 1996) for choosing between models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)}$.

BIC $\Rightarrow$ penalty equal to $F = \log n$ for each parameter used.

The overall (posterior) penalty imposed to the deviance measure will be equal to

$$(d_{\boldsymbol{\gamma}^{(k)}} - \boldsymbol{\gamma}^{(\ell)})\log n + \xi_{k\ell}.$$

3.2 Prior distributions

Prior on model parameters

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

- 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.

A cost-penalized prior on model space (1)
Preliminaries

- We propose to specify our prior model probabilities via **cost-dependent penalties** for each variable.
- We denote by c_j the cost of X_j covariate and by $\boldsymbol{c} = (c_1, c_2, \ldots, c_p)$ the vector of the costs of all variables under consideration.
- To specify this prior we define a baseline cost c_0 which is assumed to be a low acceptable cost for the collection of the data of a covariate. All the variable costs can be then written as $c_j = k_j c_0$.
- We assume that a Bayesian cost-benefit analysis is implemented using a uniform prior on model space.

A cost-penalized prior on model space (2)
The five criteria

We specify our prior distribution on $\boldsymbol{\gamma}$ to satisfy the following five criteria:

(a) The prior must be unaffected to transformations $\boldsymbol{c} \mapsto \alpha\,\boldsymbol{c}$ with $\alpha > 0$, so that conversion between time and money or different monetary units (e.g., dollars and euros) leaves the prior unchanged;

(b) the extra penalty ξ_1 for adding a variable X_j with baseline cost c_0 is zero;

(c) the extra penalty ξ_2 for adding a variable X_j with cost $c_j = \kappa\, c_0$ for some $\kappa > 1$ equals the BIC penalty of $(\kappa - 1)$ variables with cost c_0;

(d) the extra penalty ξ_3 for adding any variable X_j is greater or equal to zero; and

(e) if all the variables have the same cost, then the prior must reduce to the uniform prior on $\boldsymbol{\gamma}$.

A cost-penalized prior on model space (3)
The five criteria - interpretation

(a) ensures that the prior is invariant with respect to the manner in which cost is measured.

(b) ensures that the penalty for adding a variable X_j with baseline cost c_0 is the same as in the benefit-only analysis.

(c) ensures that the posterior model odds will still have a BIC-like behavior. The induced extra penalty will be equal to the relative difference between the cost of X_j and a variable with cost equal to c_0.

(d) ensures that the cost-benefit analysis will support more parsimonious models than the corresponding ones supported by the benefit-only analysis.

(e) requires that our prior should reproduce the benefit-only analysis if all costs are equal.

A cost-penalized prior on model space (4)
The prior

The following theorem provides the only prior that meets the above five requirements, and defines the choice of c_0.

Theorem 1. *If a prior distribution $f(\gamma)$ satisfies requirements (a-e) above, then it must be of the form*

$$f(\gamma_j) \propto \exp\left[-\frac{\gamma_j}{2}\left(\frac{c_j}{c_0}-1\right)\log n\right] \text{ for } j=1,\ldots,p, \tag{2}$$

where c_j is the marginal cost per observation for variable X_j and $c_0 = \min\{c_j, j=1,\ldots,p\}$.

For proof see Fouskakis, Ntzoufras and Draper (2008, *Annals of Applied Statistics*, to appear).

3.3 Posterior model odds

Cost-adjusted generalization of BIC

Under the above prior, if we consider the BIC-based approximation (1) then

$$-2\log PO_{k\ell} = -2\log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right) + \frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}\log n + O(1). \qquad (3)$$

where $C_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j c_j$ is the cost of model $\boldsymbol{\gamma}$.

- The penalty term $d_{\boldsymbol{\gamma}} \log n$ of model $\boldsymbol{\gamma}$ used in (1) has been replaced in the above expression by the cost-dependent penalty $c_0^{-1} C_{\boldsymbol{\gamma}} \log n$;
- Ignoring costs is equivalent to $c_j = c_0$ for all j, yielding $c_0^{-1} C_{\boldsymbol{\gamma}} = d_{\boldsymbol{\gamma}}$, the original BIC expression.
- We may interpret $\log n$ as the imposed penalty for each variable included in the model when no costs are considered.
- This baseline penalty term is inflated proportionally to the cost ratio $\frac{c_j}{c_0}$ for each X_j; for example, if the cost of a variable X_j is twice the minimum cost ($c_j = 2\, c_0$) then the imposed penalty is equivalent to adding two variables with the minimum cost.
- For all these reasons, (3) can be considered as a cost-adjusted generalization of BIC when prior model probabilities of type (2) are adopted.

3.4 Implementation and results

Implementation details

- The procedure
  1. Run RJMCMC (Green, 1995) for 100K iterations in the full model space.
  2. Eliminate non-important variables (with marginal probabilities < 0.30) forming a new reduced model space.
  3. Run RJMCMC for 100K iterations in the reduced model space to estimate posterior model odds and best models.
- Two setups:
  1. Benefit only analysis (uniform prior on model space).
  2. Cost - Benefit Analysis (cost penalized prior on model space).

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

Variable Index	Variable Name	Variable Cost	Benefit Analysis	Cost-Benefit Analysis
1	**Systolic Blood Pressure (SBP) Score**	0.50	0.99	0.99
2	**Age**	0.50	0.99	0.99
3	**Blood Urea Nitrogen**	1.50	1.00	0.99
4	**Apache II Coma Score**	2.50	1.00	
5	**Shortness of Breath Day 1**	1.00	0.97	0.79
8	**Septic Complications**	3.00	0.88	
12	**Initial Temperature**	0.50	0.98	0.96
13	Heart Rate Day 1	0.50		0.34
14	Chest Pain Day 1	0.50		0.39
15	Cardiomegaly Score	1.50	0.71	
27	Hematologic History Score	1.50	0.45	
37	Apache Respiratory Rate Score	1.00	0.95	0.32
46	Admission SBP	0.50	0.68	0.90
49	Respiratory Rate Day 1	0.50		0.81
51	Confusion Day 1	0.50		0.95
70	Apache pH Score	1.00	0.98	0.98
73	Morbid + Comorbid Score	7.50	0.96	
78	Musculoskeletal Score	1.00		0.54
	Number of variables		13	13

Reduced Model Space: Posterior Model Probabilities/Odds

Common variables in both analyses: $X_1 + X_2 + X_3 + X_5 + X_{12} + X_{70}$

Benefit-Only Analysis

k	Common Variables Within Each Analysis	Additional Variables	Model Cost	Posterior Probabilities *	PO_{1k}^{**}
1	$X_4 + X_{15} + X_{37} + X_{73}$	$+X_8 + X_{27} + X_{46}$	22.5	0.3066	1.00
2		$+X_8 + X_{27}$	22.0	0.1969	1.56
3		$+X_8$	20.5	0.1833	1.67
4		$+X_{27} + X_{46}$	19.5	0.0763	4.02
5			17.5	0.0383	8.00

Cost-Benefit Analysis

k	Common Variables Within Each Analysis	Additional Variables	Model Cost	Posterior Probabilities *	PO_{1k}^{**}
1	$X_{46} + X_{51}$	$+X_{49} + X_{78}$	7.5	0.1460	1.00
2		$+X_{14} + X_{49} + X_{78}$	7.5	0.1168	1.27
3		$+X_{13} + X_{49} + X_{78}$	7.5	0.0866	1.69
4		$+X_{13} + X_{14} + X_{49} + X_{78}$	8.0	0.0665	2.20
5		$+X_{14} + X_{49}$	7.0	0.0461	3.17
6		$+X_{49}$	6.5	0.0409	3.57
7		$+X_{37} + X_{78}$	7.5	0.0382	3.82
8		$+X_{13} + X_{14} + X_{49}$	7.5	0.0369	3.96
9		$+X_{13}$	6.5	0.0344	4.25

* above 3%. ** posterior odds of the best model within each analysis versus the current model k.

Reduced Model Space: Comparisons

Comparison of measures of fit, cost and dimensionality between the best models in the reduced model space of the benefit-only and cost-benefit analysis; percentage difference is in relation to benefit-only.

	Analysis		Difference
	Benefit-Only	Cost-Benefit	(%)
Minimum Deviance	1553.2	1635.8	+5.3
Median Deviance	1564.5	1644.8	+5.1
Cost	22.5	7.5	–66.7
Dimension	13	10	–23.1

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 $\boldsymbol{\gamma}$ with total cost larger than C, resulting to a significantly reduced model space,

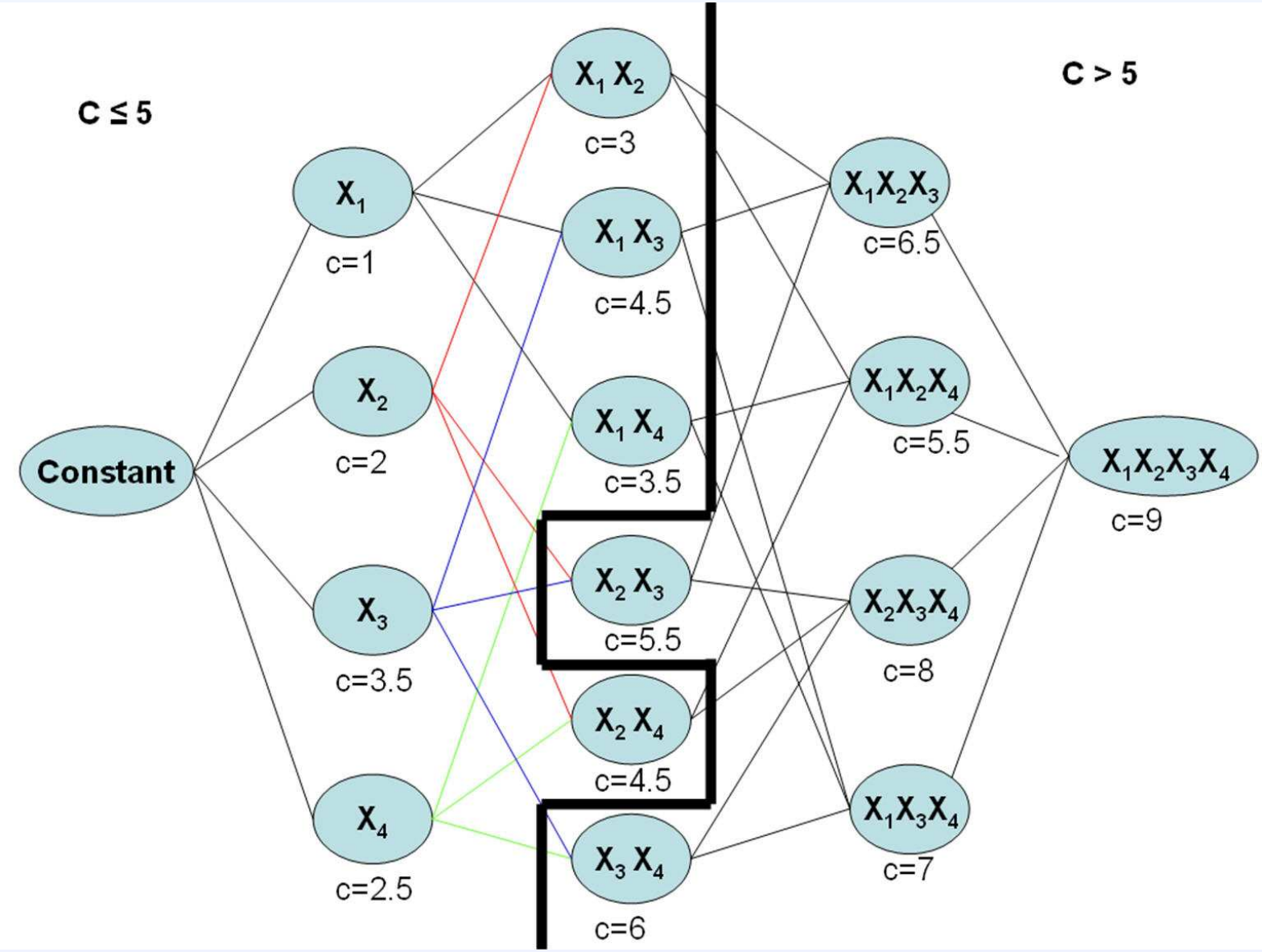
$$\mathcal{M} = \left\{ \boldsymbol{\gamma} \in \{0,1\}^p : \sum_{j=1}^{p} c_j \gamma_j \leq C \right\}.$$

- **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.

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 (**temperature**), $k = 1, 2$.
  - **1st auxiliary chain**: $t_1 > 1 \rightarrow$ increasing differences between the posterior probabilities (makes the distribution steeper allowing by this way the MCMC to move closer to locally best models).
  - **2nd auxiliary chain**: $0 < t_2 < 1 \rightarrow$ reducing differences between the posterior probabilities (makes the distribution flatter allowing by this way the MCMC to move easily across different models).
- Temperatures t_k change stochastically.
- By this way the extensive number of chains is avoided (usually from 5-10 in population based samplers).

Population RJMCMC (2)

- The **incorporation of stochastic temperatures** can be done using pseudo priors $g_k(t_k)$.
- The posterior distribution is expanded to

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma},(1)}, \boldsymbol{\gamma}_{(1)}, \boldsymbol{\beta}_{\boldsymbol{\gamma},(2)}, \boldsymbol{\gamma}_{(2)}, t_1, t_2 | \boldsymbol{y})$$

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

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

Population RJMCMC (2)

- Model indicators and parameters can be updated using RJMCMC steps.
- In Gibbs sampling, the temperature t_k is generated by

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

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

  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 temperatures t_k $f(t_k|\boldsymbol{y})$ in the sampling scheme.

- The desired posterior marginal distribution for the temperatures t_k is given by

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

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

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

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

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

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

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

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

Population RJMCMC (3)

Our algorithm can be summarised as follows:

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

$$A = \left[\frac{f\left(\boldsymbol{y}|\boldsymbol{\beta}'_{\boldsymbol{\gamma},(k)}, \boldsymbol{\gamma}'_{(k)}\right) f\left(\boldsymbol{\beta}'_{\boldsymbol{\gamma},(k)}|\boldsymbol{\gamma}'_{(k)}\right) f\left(\boldsymbol{\gamma}'_{(k)}\right)}{f\left(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}, \boldsymbol{\gamma}_{(k)}\right) f\left(\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}|\boldsymbol{\gamma}_{(k)}\right) f\left(\boldsymbol{\gamma}_{(k)}\right)}\right]^{t_k} \frac{q_{j,k}(\beta_{j,(k)})^{\gamma_{j,(k)}}}{q_{j,k}(\beta'_{j,(k)})^{1-\gamma_{j,(k)}}}. \tag{4}$$

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 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 $\alpha = \min\{1, A\}$, where

$$A = \left[\frac{f\left(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}, \boldsymbol{\gamma}_{(k)}\right) f\left(\boldsymbol{\beta}_{\boldsymbol{\gamma},(k)}|\boldsymbol{\gamma}'_{(k)}\right) f\left(\boldsymbol{\gamma}'_{(k)}\right)}{f\left(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\right) f\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) f(\boldsymbol{\gamma})}\right]^{1-t_k}. \tag{5}$$

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.

4.2 Prior Distributions

Same prior on model parameters as in the Cost - Benefit Analysis and a uniform prior on cost restricted model space, i.e.

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

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 $f(\gamma_j = 1|\boldsymbol{y})$

Variables with marginal posterior probabilities $f(\gamma_j = 1|\boldsymbol{y})$ *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 Posterior Prob.	1st Run PO^*_{1k}	2nd Run Posterior Prob.	2nd Run 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 Posterior Prob.	1st Run PO^*_{1k}	2nd Run Posterior Prob.	2nd Run 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

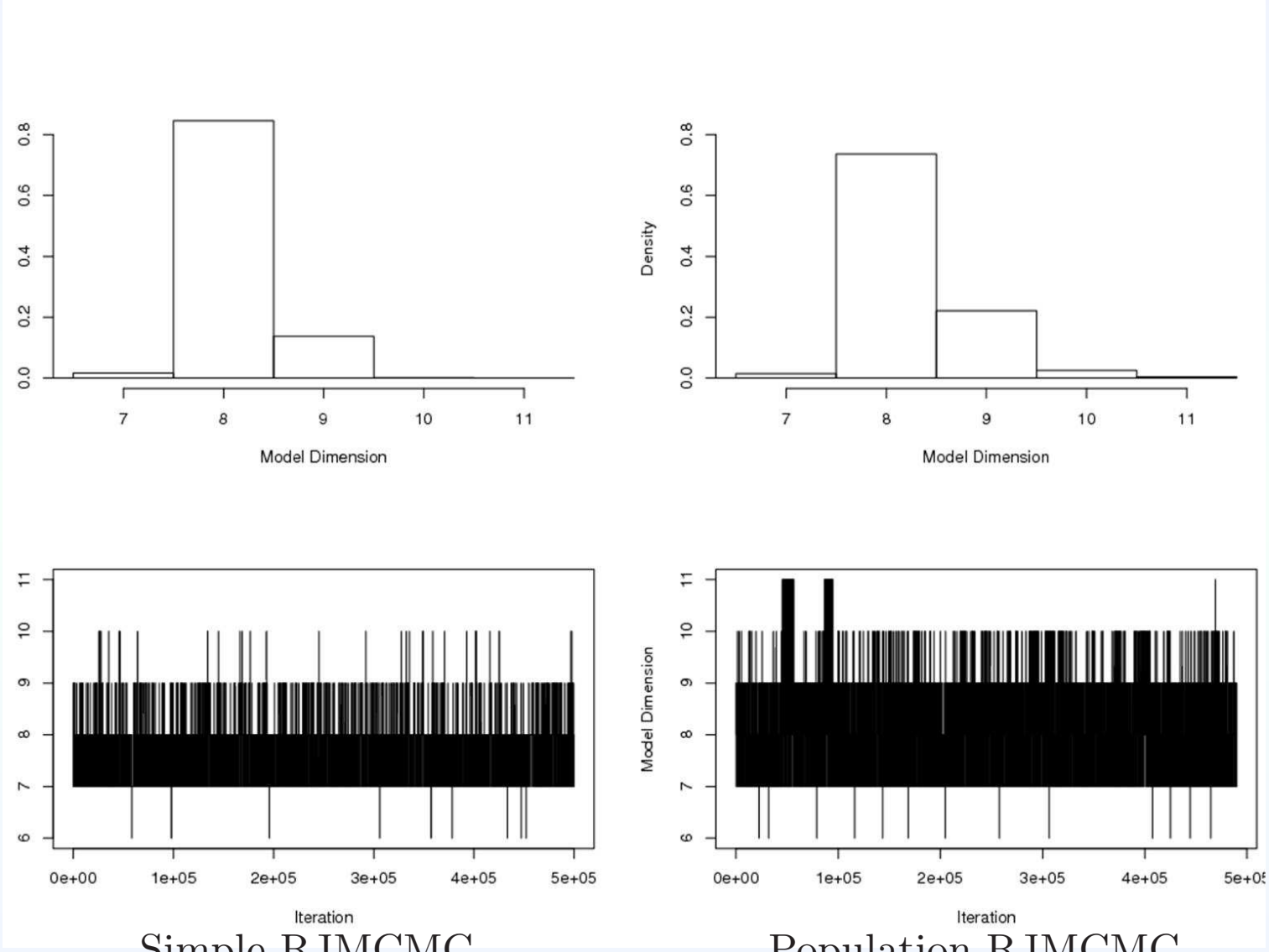
RJMCMC			MCSEs (%)			
Type	Run	Iterations	m_1	m_2	m_3	m_4
P	1	500K	1.2	0.5	0.9	0.7
P	2	500K	1.5	0.4	1.0	0.7
P	1	200K	1.9	0.8	1.1	1.2
P	2	200K	1.6	1.0	1.1	0.9
P	1	100K	2.5	1.2	1.7	1.5
P	2	100K	2.7	0.9	1.6	1.2
S	1	500K	4.2	1.3	3.2	0.0
S	2	500K	4.2	1.7	3.6	0.0
S	1	1,500K	2.9	1.1	2.1	1.0
S	2	1,500K	3.1	0.9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

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

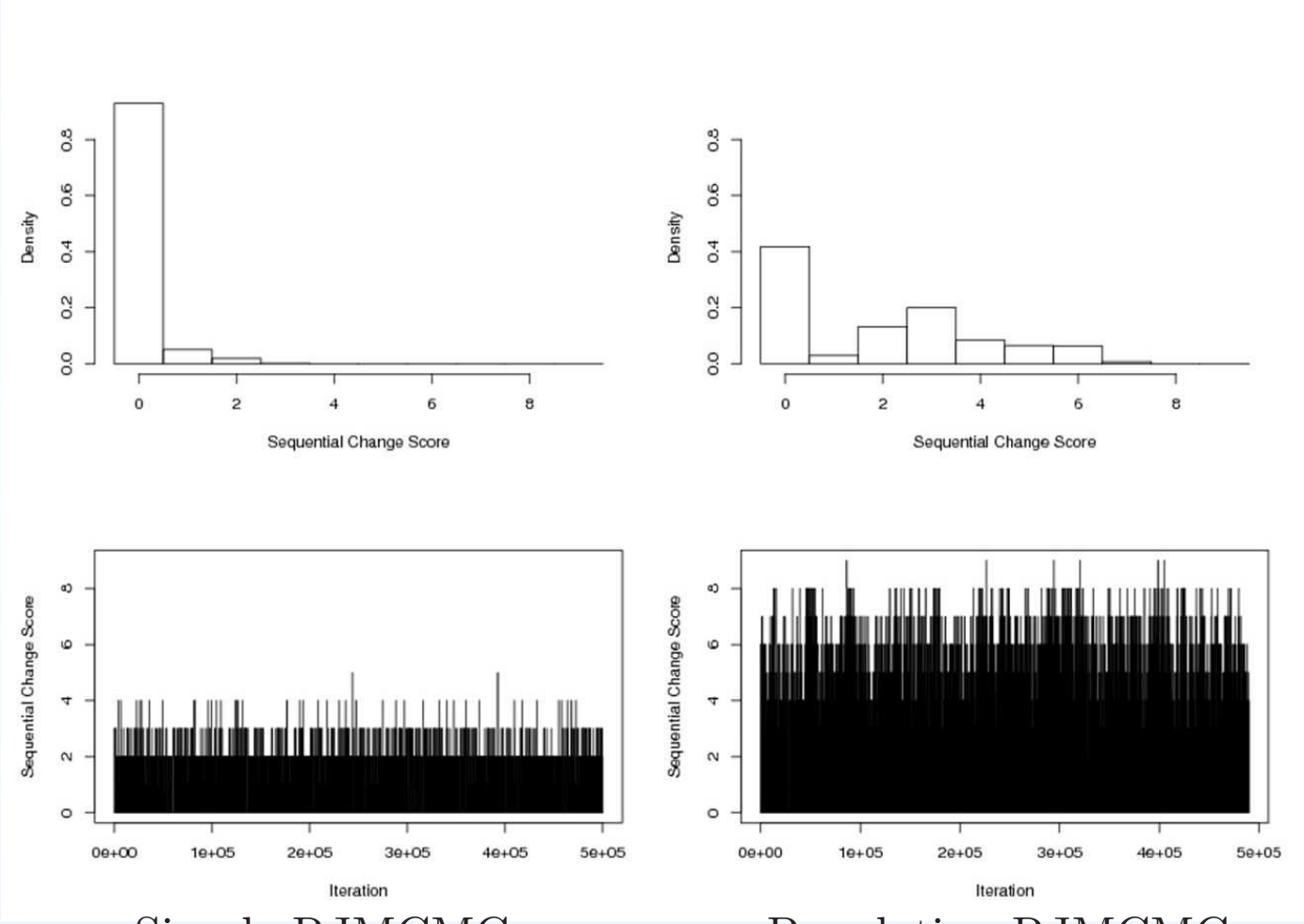
Figure 1: Density and time series plots of model dimension.



Simple RJMCMC
1,500K runs
(thinned by a factor of 3)

Population RJMCMC
500K runs

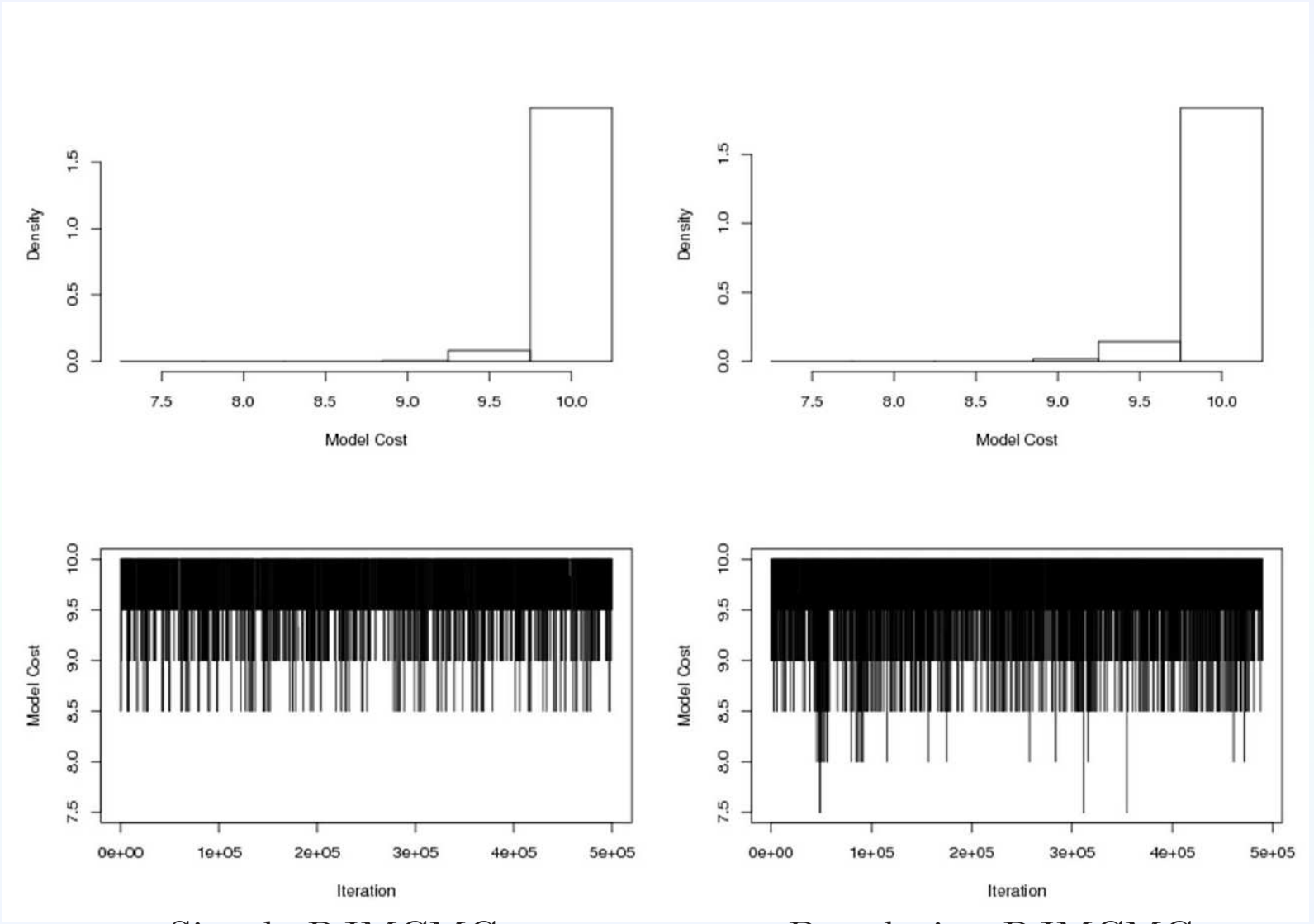
Figure 2: Density and time series plots of sequential change score (the number of variables in the model at iteration $(t+1)$ that are different from those in the model at iteration t)



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 - Benefit Analysis**:

  We have specified cost adjusted prior model probabilities resulting in posterior model odds that can be approximated by a cost adjusted BIC measure. The prior specification was achieved via the definition of five criteria that ensure the plausibility and consistency of our prior.

  Posterior analysis using the proposed prior setup achieves dramatic gains in cost and noticeable improvement in model simplicity at the price of a small loss in predictive accuracy, when compared to the results of a more traditional benefit-only analysis.

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

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

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

Authors' related work

Draper D, Fouskakis D (2000). A case study of stochastic optimization in health policy: problem formulation and preliminary results. *Journal of Global Optimization*, **18**, 399–416.

Fouskakis D, Draper D (2002). Stochastic optimization: a review. *International Statistical Review*, **70**, 315–349.

Fouskakis D, Draper D (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. *Journal of the American Statistical Association*, **103**, forthcoming.

Fouskakis D, Ntzoufras I, Draper D (2008a). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics* (to appear).

Fouskakis D, Ntzoufras I, Draper D (2008b). Population Based Reversible Jump MCMC for Bayesian Variable Selection and Evaluation Under Cost Limit Restrictions. *Journal of the Royal Statistical Society C (Applied Statistics)* (to appear).

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.