Incorporating cost in the Bayesian inference for the assessment of health care quality

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 Technica 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 will be available at: http://stat-athens.aueb.gr/~jbn/current.htm

Synopsis

1. Motivation - Presentation of the Problem and the Data.
2. Model Specification.
3. Cost - Benefit Analysis.
4. Cost Restriction - Benefit Analysis.
5. Discussion.

1 Motivation - Presentation of the Problem and the Data.

How to measure hospital quality of care?

- Indirect method: **input-output** approach (also called **league table** quality assessment) — hospital outcomes (e.g., death within 30 days of admission) compared *after adjusting for differences in inputs* (sickness at admission).
- **Cost-effective** measurement of admission sickness crucial to this approach.

Data

- Available **inputs** to sickness scale: 80–100 variables (e.g., blood urea nitrogen, coma score).
- Outcome = 30-day death (binary).

Usual method of Analysis

Logistic regression using **frequentist variable-selection methods** to find parsimonious and clinically reasonable subset.

Data in this study

- Quality of hospital care US study by conducted RAND Corporation.
- Sample: $n = 2532$ pneumonia patients in the late 1980s (Kahn, *et al.* , 1990)
- Logistic regression was used to reduce the initial list of $p = 83$ available sickness indicators for pneumonia down to a 14 predictors (Keeler, *et al.* , 1990).

- This approach is sub-optimal: it does not consider **differences in cost of data collection** among available predictors.
- Cost is measured in data collection time ranging from 30 seconds to 10 minutes of abstraction time per variable; Data collectors payment is roughly at 20$/hour.
- Weighing data-collection cost against accuracy of prediction $\rightarrow$ large **variable selection problem** (when $p = 83$ we need to compare $\boxed{2^p \approx 9.7 \cdot 10^{24}}$ subsets of sickness variables).
- For a decision theoretic approach of the same problem see Fouskakis and Draper (2007).

The 14-Variable Rand Pneumonia Scale

Admission sickness scale created by Rand for pneumonia patients contained $p = 14$ variables, chosen to **optimize predictive accuracy** subject to informal parsimony constraint (CHF = congestive heart failure).

Variable	Cost c_j (minutes)	Variable	Cost c_j (minutes)
Total **APACHE II** score	10.00	**Age**	0.50
Systolic **blood pressure** score (2-point scale)	0.50	Chest X-ray **CHF score** (3-point scale)	2.50
Blood urea nitrogen (**BUN**)	1.50	APACHE II **coma score** (3-point scale)	2.50
Serum albumin (3-point scale)	1.50	**Shortness of breath** (yes, no)	1.00
Respiratory distress (yes, no)	1.00	**Septic complications** (yes, no)	3.00
Prior **respiratory failure** (yes, no)	2.00	**Recently hospitalized** (yes, no)	2.00
Ambulatory score (3-point scale)	2.50	**Temperature**	0.50

2 Model Specification

- Logistic regression model with $Y_i = 1$ if patient i dies.
- 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

$$(Y_i \mid \boldsymbol{\gamma}) \overset{indep}{\sim} Bernoulli(p_i(\boldsymbol{\gamma})), \tag{1}$$

$$\eta_i(\boldsymbol{\gamma}) = \log\left(\frac{p_i(\boldsymbol{\gamma})}{1 - p_i(\boldsymbol{\gamma})}\right) = \sum_{j=0}^{p} \beta_j \gamma_j X_{ij}, \tag{2}$$

$$\tag{3}$$

Two different approaches

1. Cost-benefit analysis (incorporate cost in the analysis)
2. Cost Restricted benefit analysis (impose a cost limit on the use of variables)

3 Cost-Benefit Analysis

The aim is to identify well fitted models after taking into account the cost of each variable (Fouskakis, *et al.* , 2007).

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.

Prior Distributions

Prior on model parameters (see Ntzoufras *et al.* , 2003)

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

Use a cost penalized prior for variable inclusion indicators

$$f(\gamma_j) \propto \exp\left(\frac{\gamma_j}{2}\ \frac{c_0 - c_j}{c_0}\log n\right) \quad \text{for} \quad j = 1, \ldots, p\ . \tag{5}$$

When comparing models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)} \Rightarrow$ penalty imposed to the log-likelihood ratio:

$$-2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} = \sum_{j=1}^{p}\left(\gamma_j^{(k)} - \gamma_j^{(\ell)}\right)\frac{c_j - c_0}{c_0}\log n\ .$$

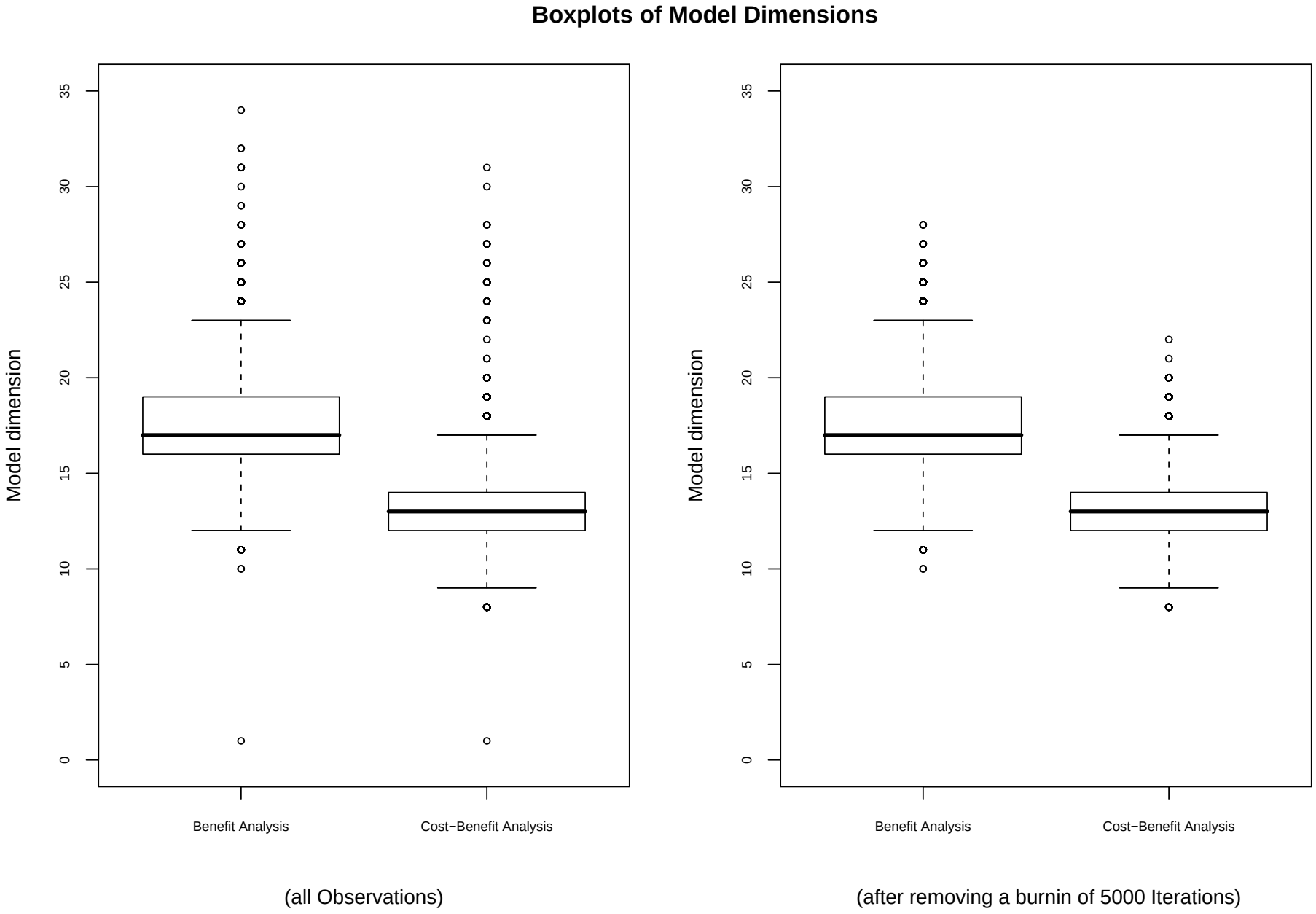
- c_j : cost per observation for X_j variable.
- c_0 : baseline cost (default choice: $c_0 = \min\{c_j\}\ \forall\ j = 1, \ldots, p$).
- Indifference concerning the cost $\Rightarrow c_j = c_0$ for $j = 1, \ldots, p \Rightarrow$ uniform prior on model space ($f(\boldsymbol{\gamma}) \propto 1$) $\Rightarrow$ Posterior model odds = Bayes factor.

Implementation and Results

- Run RJMCMC (Green, 1995) for 100K iterations in the full model space.
- Eliminate unimportant variables (with marginal probabilities < 0.30) forming a new reduced model space.
- 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).

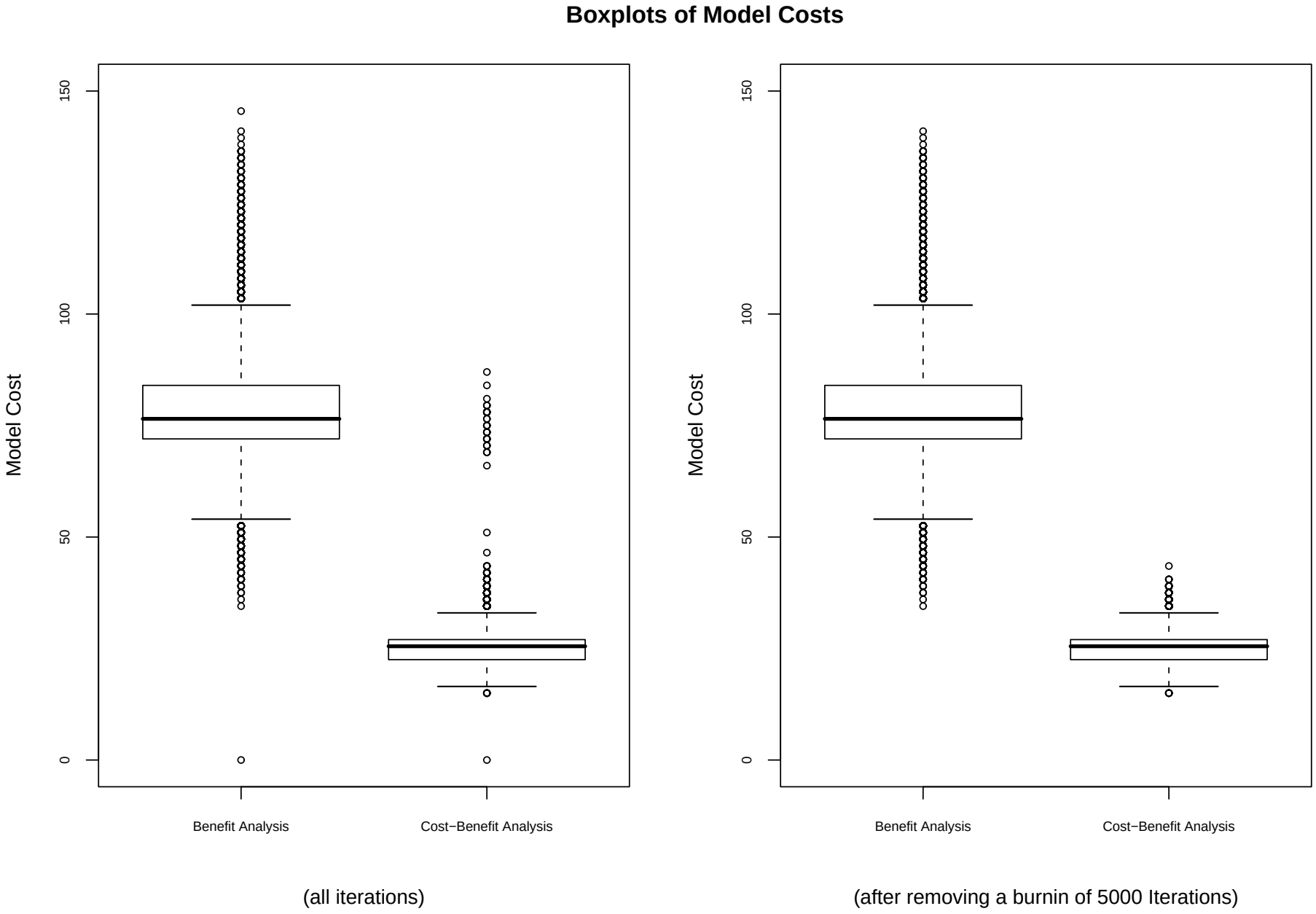
Preliminary Results: Marginal Probabilities $f(\gamma_j = 1|\boldsymbol{y})$

Variable Index	Variable Name	Costs (minutes)	Benefit Analysis	Cost-Benefit Analysis
1	Systolic Blood Pressure 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	1.00	0.97	0.79
8	Septic Complications	3.00	0.88	
12	Temperature	0.50	0.98	0.96
13	Heart Rate	0.50		0.34
14	Chest Pain	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	0.50		0.81
51	Confusion	0.50		0.95
70	Apache PH Score	1.00	0.98	0.98
73	Morbid + Comorbid	7.50	0.96	
78	Musculoskeletal Score	1.00		0.54



(all Observations)

(after removing a burnin of 5000 Iterations)



(all iterations)

(after removing a burnin of 5000 Iterations)

Reduced Model Space: Marginal Probabilities $f(\gamma_j = 1|\boldsymbol{y})$

Variable Index	Variable Name	Costs (minutes)	Benefit Analysis	Cost-Benefit Analysis
1	Systolic Blood Pressure Score	0.50	1.00	0.99
2	Age	0.50	1.00	1.00
3	Blood Urea Nitrogen	1.50	1.00	1.00
4	Apache II Coma score	2.50	1.00	
5	Shortness of Breath	1.00	0.97	0.89
8	Septic Complications	3.00	0.89	
12	Temperature	0.50	0.99	0.95
13	Heart Rate	0.50		0.37
14	Chest Pain	0.50		0.45
15	Cardiomegaly Score	1.50	0.90	
27	Hematologic History Score	1.50	0.66	
37	Apache Respiratory Rate Score	1.00	1.00	0.28
46	Admission SBP	0.50	0.63	0.94
49	Respiratory Rate	0.50		0.84
51	Confusion	0.50		1.00
70	Apache PH Score	1.00	0.99	1.00
73	Morbid + Comorbid	7.50	1.00	
78	Musculoskeletal Score	1.00		0.71

Reduced Model Space: Posterior Model Probabilities/Odds

Set-up	k	Common Variables in all set-ups	Common Variables within set-up	Additional Variables	Post. Prob.	PO^*_{1k}
Benefit	1	X1+X2+X3+X5+X12+X70	+X4+X15+X37+X73	+X8 +X27+X46	0.3066	1.00
Analysis	2			+X8 +X27	0.1969	1.56
	3			+X8	0.1833	1.67
	4			+X27+X46	0.0763	4.02
	5				0.0383	8.00
Cost	1	X1+X2+X3+X5+X12+X70	+X46+X51	+X49+X78	0.1460	1.00
Benefit	2			+X14 +X49+X78	0.1168	1.27
Analysis	3			+X13 +X49+X78	0.0866	1.69
	4			+X13+X14 +X49+X78	0.0665	2.20
	5			+X14 +X49	0.0461	3.17
	6			+X49	0.0409	3.57
	7			+X37 +X78	0.0382	3.82
	8			+X13+X14 +X49	0.0369	3.96
	9			+X13	0.0344	4.25

* Posterior odds of the best model within each set-up versus the current model k

Reduced Model Space: Comparisons

Comparison of measures of fit, cost and dimensionality between the visited 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	(%)
Min Deviance	1553.2	1616.1	+4.1
Median Deviance	1572.0	1643.8	+4.6
Median Cost	22.0	7.5	–65.9
Median Dimension	13	11	–15.4

4 Cost Restriction - Benefit Analysis

- Usually, a cost limit is imposed by the project budget.
- Hence, the search should be conducted only among models whose cost does not exceed the 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,

$$\mathcal{M} = \{\boldsymbol{\gamma} \in \{0,1\}^p : \sum_{i=1}^{p} c_i \gamma_i \leq C\}.$$

- AIM: estimate posterior model probabilities in the cost restricted model space.
- PROBLEM: due to the cost limit, model space areas of local maximum may exist.
- 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).
- We have combined ideas from the population-based MCMC (Jasra, *et al.* , 2007) and simulated tempering algorithms (Geyer and Thompson, 1995)).

Population based MCMC

- Use 3 chains: The actual one and two auxiliary ones.
  - Auxiliary chains are equal to the posterior distributions raised in a power t_k called temperature.
  - 1st auxiliary chain: $t_k > 1$ increasing differences between the posterior probabilities (makes the distribution steeper allowing by this way the MCMC to move closser to locally best models).
  - 2nd auxiliary chain: $0 < t_k < 1$ 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.

Prior Distributions

Same prior on model parameters as before

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = N\left[\mathbf{0}, 4n\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\right)^{-1}\right] \tag{6}$$

and a uniform prior on cost restricted model space, i.e.

$$f(\boldsymbol{\gamma}) \propto I(\boldsymbol{\gamma} \in \mathcal{M} : c(\boldsymbol{\gamma}) = \sum_{j=1}^{p} \gamma_j c_j \leq C), \tag{7}$$

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

Implementation and Results

- COST LIMIT: $C = 10$ minutes of abstraction time.
- Run Population RJMCMC for 100K iterations twice in the full model space.
- Eliminate unimportant variables (with marginal probabilities < 0.30 in any run) forming a new reduced model space.
- Run population RJMCMC in the reduced space (twice).
- Compare results and performance of population RJMCMC with simple RJMCMC.

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; costs are expressed in minutes of abstraction time.

Variable			Marginal Posterior Probabilities	
			First Run	Second Run
Index	Name	Cost	Analysis	Analysis
1	Systolic Blood Pressure 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	1.00	0.92	0.80
6	Serum Albumin Score	1.50	0.40	0.55
12	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	0.50	0.35	0.25
51	Confusion	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

Population RJMCMC - 500K iterations

k	m	Common Variables	Additional Variables	First Run Model Prob.	First Run Post. PO_{1k}	Second Run Model Prob.	Second Run Post. PO_{1k}
1	m_1	$X_1 + X_{12} + X_{37}$	$+X_3+X_5 \quad +X_{62}$	0.4872	1.00	0.4879	1.00
2	m_2		$+X_5 \quad +X_{46} \; +X_{62} \; +X_{70}$	0.1202	4.05	0.1052	4.63
3	m_3		$+X_3 \quad +X_{62} \; +X_{70}$	0.0894	5.45	0.0982	4.97
4	m_4		$+X_3+X_5+X_6 \quad +X_{70}$	0.0344	14.16	0.0498	9.80

Simple RJMCMC - 500K iterations

k	m	Com. Vars	Additional Variables	1st Run Model Prob.	1st Run Post. PO_{1k}	2nd Run Model Prob.	2nd Run Post. PO_{1k}
1	m_1	X_{62}	$+X_1+X_3+X_5+X_{12}+X_{37}$	0.6129	1.00	0.5952	1.00
2	m_3		$+X_1+X_3 \quad +X_{12}+X_{37} \quad +X_{70}$	0.0828	7.40	0.1214	4.90
3	m_2		$+X_1 \quad +X_5+X_{12}+X_{37}+X_{46} \quad +X_{70}$	0.0762	8.04	0.1074	5.54
4	m_5		$+X_3+X_5 \quad +X_{46} \; +X_{49} \; +X_{70}$	0.0457	13.41	< 0.03	> 19.9
5	m_6		$+X_1+X_3+X_5 \quad +X_{49} \; +X_{70}$	0.0337	18.19	< 0.03	> 19.9

Common variables in all analyses: $X_2 + X_4$ *All models appearing in the table have total cost 10 min (cost limit).*

Reduced Model Space: Monte Carlo Errors

Monte Carlo Errors (%)						
RJMCMC Type	Run	Iterations	m_1	m_2	m_3	m_4
POP.	1	500K	1.2	0.5	0.9	0.7
POP.	2	500K	1.5	0.4	1.0	0.7
POP.	1	200K	1.9	0.8	1.1	1.2
POP.	2	200K	1.6	1.0	1.1	0.9
POP.	1	100K	2.5	1.2	1.7	1.5
POP.	2	100K	2.7	0.9	1.6	1.2
SIMPLE	1	500K	4.2	1.3	3.2	0.0
SIMPLE	2	500K	4.2	1.7	3.6	0.0
Relative Comparisons						
SIMPLE vs. POP. (First Run)		500K	3.5	2.8	3.6	0.0
		200K	2.2	1.8	2.9	0.0
		100K	1.7	1.2	1.9	0.0
SIMPLE vs. POP. (Second Run)		500K	2.8	3.4	3.6	0.0
		200K	2.6	1.7	3.3	0.0
		100K	1.6	1.9	2.3	0.0

5 Discussion

- Cost - Benefit Analysis:

  The resulting models achieve 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.

  Bayesian model averaging (accounting also for the cost) is feasible.

- Cost Restriction - Benefit Analysis:

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

References

Draper D, Kahn K, Reinisch E, Sherwood M, Carney M, Kosecoff J, Keeler E, Rogers W, Savitt H, Allen H, Well K, Reboussin D, Brook R (1990). Studying the effects of the DRG-based Prospective Payment System on Quality of Care: Design, Sampling, and fieldwork. *Journal of the American Medical Association*, **264**, 1956–1961.

Fouskakis, D. and Draper, D. (2007). Stochastic optimization methods for cost-effective quality assessment in health. (submitted).

Fouskakis, D., Ntzoufras, I. and Draper, D. (2007). Bayesian variable selection using a cost-penalized approach, with application to cost-effective measurement of quality of health care. (submitted).

Geyer, C.J. and Thomson, E.A. (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, D.A. and Holmes, C.C. (2007). Population-based reversible jump MCMC. *Biometrika*. (to appear).

Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990). The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953–1955 (with editorial comment, 1995–1997).

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.