Bayesian Modeling, Inference, Prediction and Decision-Making

7: A Practical Example of Bayesian Decision Theory

David Draper (draper@ams.ucsc.edu)

Department of Applied Mathematics and Statistics
University of California, Santa Cruz (USA)

Short Course (sponsored by eBay and google)

10 Fridays, 11 Jan-22 Mar 2013 (except 25 Jan)

course web page

www.ams.ucsc.edu/∼draper/eBay-2013.html

The Problem

Variable selection (choosing the **"best" subset of predictors**) in **generalized linear models** is an **old problem**, dating back at least to the **1960s**, and **many methods** have been proposed to **try to solve it**; but virtually all of them **ignore** an **aspect** of the problem that can be **important**: the **cost of data collection of the predictors**.

$\boxed{\textbf{Case study.}}$ (Fouskakis and Draper, *JASA*, 2008; Fouskakis, Ntzoufras and Draper (FND), submitted, 2007a, 2007b). In the field of **quality of health care measurement**, patient **sickness at admission** is traditionally assessed by using **logistic regression** of **mortality within 30 days of admission** on a fairly large number of **sickness indicators** (on the order of **100**) to construct a sickness scale, employing standard **variable selection methods** (e.g., **backward selection** from a model with all predictors) to find an **"optimal"** subset of **10–20** indicators.

Such **"benefit-only"** methods ignore the considerable **differences** among the sickness indicators in **cost of data collection**, an issue that's **crucial** when admission sickness is used to drive programs (now implemented or

Choosing Utility Function (continued)

under consideration in several countries, including the U.S. and U.K.) that attempt to **identify substandard hospitals** by comparing **observed** and **expected mortality rates (given admission sickness)**.

When both **data-collection cost** and **accuracy of prediction** of 30-day mortality are considered, a large **variable-selection problem** arises in which **costly variables** that **do not predict well enough** should be **omitted** from the final scale.

There are **two main ways** to solve this problem — you can (a) put **cost** and **predictive accuracy** on the **same scale** and **optimize**, or (b) **maximize** the latter subject to a **bound** on the former — leading to **three methods**:

(1) a **decision-theoretic cost-benefit approach** based on **maximizing expected utility** (Fouskakis and Draper, 2008),

(2) an **alternative cost-benefit approach** based on **posterior model odds** (FND, 2007a), and

(3) a **cost-restriction-benefit analysis** that **maximizes predictive accuracy** subject to a **bound on cost** (FND, 2007b).

The Data

Data (Kahn et al., *JAMA*, 1990): $p = 83$ **sickness indicators** gathered on **representative sample** of $n = 2,532$ elderly American patients hospitalized in the period 1980–86 with **pneumonia**; original RAND **benefit-only scale** based on **subset** of 14 predictors:

Variable	Cost (U.S.$)	Correlation	Good?
Total APACHE II score (36-point scale)	3.33	0.39	
Age	0.50	0.17	$*$
Systolic blood pressure score (2-point scale)	0.17	0.29	$**$
Chest X-ray congestive heart failure score (3-point scale)	0.83	0.10	
Blood urea nitrogen	0.50	0.32	$**$
APACHE II coma score (3-point scale)	0.83	0.35	$**$
Serum albumin (3-point scale)	0.50	0.20	$*$
Shortness of breath (yes, no)	0.33	0.13	$**$
Respiratory distress (yes, no)	0.33	0.18	$*$
Septic complications (yes, no)	1.00	0.06	
Prior respiratory failure (yes, no)	0.67	0.08	
Recently hospitalized (yes, no)	0.67	0.14	
Ambulatory score (3-point scale)	0.83	0.22	
Temperature	0.17	-0.16	$*$

Decision-Theoretic Cost-Benefit Approach

$\boxed{\textbf{Approach (1)}}$ (decision-theoretic cost-benefit). **Problem formulation:** Suppose (a) the 30–day **mortality outcome** y_i and data on p **sickness indicators** $(x_{i1}, \ldots, X_{ip})$ have been collected on n individuals sampled exchangeably from a **population** $\mathcal{P}$ of patients with a given disease, and (b) the goal is to **predict** the death outcome for n^* **new patients** who will in the future be sampled exchangeably from $\mathcal{P}$, (c) on the basis of some or all of the predictors $X_{\cdot j}$, when (d) the **marginal costs of data collection** per patient $c_1, \ldots, c_p$ for the $X_{\cdot j}$ **vary considerably**.

What is the **best subset** of the $X_{\cdot j}$ to choose, if a **fixed amount of money** is available for this task and you're **rewarded** based on the **quality** of your predictions?

Since data on **future patients** are **not available**, we use a **cross-validation** approach in which (i) a random subset of n_M observations is drawn for creation of the mortality predictions (the **modeling** subsample) and (ii) the quality of those predictions is assessed on the remaining $n_V = (n - n_M)$ observations (the **validation** subsample, which serves as a proxy for future patients).

Utility Elicitation

Here **utility** is quantified in **monetary terms**, so that **data collection** part of **utility function** is simply **negative** of **total amount of money** required to gather data on specified predictor subset (**manual data abstraction** from hardcopy patient charts will gradually be replaced by **electronic medical records**, but still widely used in **quality of care studies**).

Letting $I_j = 1$ if $X_{\cdot j}$ is included in a given model (and 0 otherwise), the **data-collection utility** associated with subset $I = (I_1, \ldots, I_p)$ for patients in the **validation subsample** is

$$U_D(I) = -n_V \sum_{j=1}^{p} c_j I_j, \qquad (1)$$

where c_j is the **marginal cost per patient of data abstraction** for variable j (the second column in the table above gave examples of these marginal costs).

To measure the **accuracy** of a model's predictions, a metric is needed that quantifies the **discrepancy** between the actual and predicted values, and in this problem **the metric must come out in monetary terms** on a scale comparable to that employed with the data-collection utility.

Utility Elicitation (continued)

In the setting of this problem the outcomes Y_i are **binary death indicators** and the **predicted values** $\hat{p}_i$, based on statistical modeling, take the form of **estimated death probabilities**.

We use an approach to the comparison of **actual** and **predicted** values that involves **dichotomizing** the $\hat{p}_i$ with respect to a **cutoff**, to mimic the decision-making reality that **actions** taken on the basis of observed-versus-expected quality assessment will have an **all-or-nothing character** at the hospital level (for example, regulators must decide either to subject or not subject a given hospital to a more detailed, more expensive quality audit based on **process criteria**).

In the first step of our approach, given a particular **predictor subset** I, we fit a **logistic regression model** to the **modeling** subsample M and apply this model to **validation** subsample V to create predicted death probabilities $\hat{p}_i^I$.

In more detail, letting $Y_i = 1$ if patient i dies and 0 otherwise, and taking $X_{i1}, \ldots, X_{ik}$ to be the k **sickness predictors** for this patient under model I, the usual **sampling model** which underlies logistic regression in this case is

Utility Elicitation (continued)

$$
\begin{aligned}
(Y_i \mid p_i^I) &\overset{\text{indep}}{\sim} \text{Bernoulli}(p_i^I), \\
\log\left(\frac{p_i^I}{1-p_i^I}\right) &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.
\end{aligned}
\tag{2}
$$

We use **maximum likelihood** to fit this model (as a computationally efficient approximation to Bayesian fitting with relatively diffuse priors), obtaining a vector $\hat{\beta}$ of estimated logistic regression coefficients, from which the **predicted death probabilities** for the patients in subsample V are as usual given by

$$
\hat{p}_i^I = \left[1 + \exp\left(- \sum_{j=0}^{k} \hat{\beta}_j X_{ij} \right) \right]^{-1}, \tag{3}
$$

where $X_{i0} = 1$ ($\hat{p}_i^I$ may be thought of as the **sickness score** for patient i under model I).

In the second step of our approach we **classify** patient i in the validation subsample as **predicted dead or alive** according to whether $\hat{p}_i^I$ exceeds or falls short of a **cutoff** p^*, which is chosen — by searching on a discrete grid from 0.01 to 0.99 by steps of 0.01 — to **maximize the predictive accuracy** of model I.

Utility Elicitation (continued)

We then cross-tabulate actual versus predicted death status in a 2×2 **contingency table**, **rewarding** and **penalizing** model I according to the numbers of patients in the **validation sample** which fall into the cells of the right-hand part of the following table.

		Rewards and Penalties		Counts	
		Predicted		Predicted	
		Died	Lived	Died	Lived
Actual	Died	C_{11}	C_{12}	n_{11}	n_{12}
	Lived	C_{21}	C_{22}	n_{21}	n_{22}

The left-hand part of this table records the **rewards and penalties** in US$.

The **predictive utility** of model I is then

$$U_P(I) = \sum_{l=1}^{2} \sum_{m=1}^{2} C_{lm} \, n_{lm} . \tag{4}$$

See Fouskakis-Draper (2008) for details on **eliciting** the **utility values** C_{lm}.

Utility Elicitation (continued)

The idea was (1) to draw **correspondence** between the above 2×2 table and another such table cross-tabulating **true hospital status** (bad, good) against **action taken** (process audit, no such audit), (2) to elicit C_{21} (cost of subjecting a **good** hospital to an **unnecessary** audit) from health policy experts in the U.S. and U.K., and (3) to elicit **ratios** relating the other C_{lm} to C_{21}.

The result was $(C_{11}, C_{12}, C_{21}, C_{22}) =$ **\$(34.8, –139.2, –69.6, 8.7)**.

The findings in Fouskakis and Draper (2008) use these values; Draper and Fouskakis (2000) present a **sensitivity analysis** on the choice of the C_{lm} which demonstrates **broad stability** of the findings when the utility values mentioned above are **perturbed** in reasonable ways.

With the C_{lm} in hand, the **overall expected utility function** to be maximized over I is then simply

$$E\left[U(I)\right] = E\left[U_D(I) + U_P(I)\right], \tag{5}$$

where this expectation is over **all possible cross-validation splits** of the data.

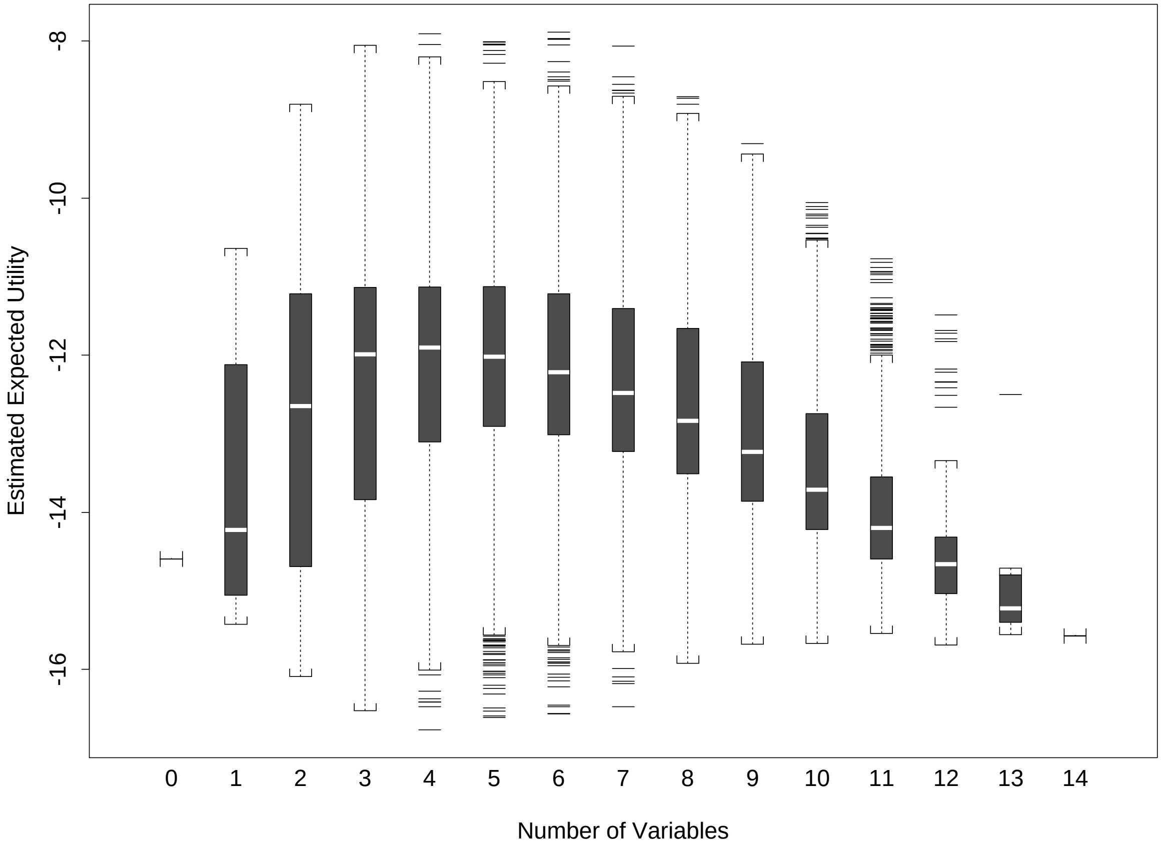
Results

The number of possible cross-validation splits is **far too large** to evaluate the expectation in (5) directly; in practice we therefore use **Monte Carlo methods** to evaluate it, **averaging** over N random modeling and validation **splits**.

$\boxed{\text{Results.}}$ We explored this approach in **two settings**:

- a **Small World** created by focusing only on the $p = 14$ variables in the **original RAND scale** ($2^{14} = 16,384$ is a **small enough number of possible models** to do **brute-force enumeration** of the estimated expected utility of all models), and
- the **Big World** defined by all $p = 83$ available predictors ($2^{83} \doteq 10^{25}$ is **far too large** for brute-force enumeration; we compared a variety of **stochastic optimization methods** — including **simulated annealing**, **genetic algorithms**, and **tabu search** — on their ability to find **good variable subsets**).

Results: Small World



The **20 best models** included the **same three variables** 18 or more times out of 20, and never included six other variables; the **five best models** were minor variations on each other, and included **4–6 variables** (last column in table on page 4).

Approach (2)

The best models **save almost $8 per patient** over the full 14-variable model; this would amount to **significant savings** if the observed-versus-expected assessment method were **applied widely**.

Approach (2) (alternative cost-benefit) **Maximizing expected utility**, as in Approach (1) above, is a natural Bayesian way forward in this problem, but (a) the elicitation process was **complicated** and **difficult** and (b) the **utility structure** we examine is only one of a number of plausible alternatives, with utility framed from **only one point of view**; the broader question for a decision-theoretic approach is **whose utility should drive the problem formulation**.

It is well known (e.g., Arrow, 1963; Weerahandi and Zidek, 1981) that **Bayesian decision theory** can be **problematic** when used **normatively** for **group decision-making**, because of **conflicts** in **preferences** among members of the group; in the context of the problem addressed here, it can be **difficult** to identify a **utility structure acceptable to all stakeholders** (including patients, doctors, hospitals, citizen watchdog groups, and state and federal regulatory agencies) in the quality-of-care-assessment process.

Approach (2) (continued)

As an **alternative**, in Approach (2) we propose a **prior distribution** that accounts for the **cost** of each variable and results in a set of **posterior model probabilities** which correspond to a **generalized cost-adjusted version of the Bayesian information criterion** (BIC).

This provides a **principled approach** to performing a **cost-benefit trade-off** that **avoids ambiguities** in identification of an **appropriate utility structure**; we reason as follows.

(1) With (a) $\boldsymbol{\gamma}^{(k)}$ as a **binary vector** specifying which variables are **included** in model k, (b) $f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})$ as the **log likelihood** of model k evaluated at the MLE $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}$ of model k's parameter vector $\beta_{\boldsymbol{\gamma}^{(k)}}$, and (c) $d_{\boldsymbol{\gamma}^{(k)}} = \sum_{j=0}^{p} \gamma_j$ as the **dimension** of model k, the usual $O(1)$ approximation to the **log posterior model odds** is

$$-2 \log PO_{k\ell} = -2 \log \left[\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})} \right] + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}} \right) \log n - 2 \log \frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(1) \qquad (6)$$

Approach (2) (continued)

$$= \quad BIC_{k\ell} - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(1),$$

where $BIC_{k\ell}$ is the **Bayesian Information Criterion** for choosing between models $\boldsymbol{\gamma}^{(k)}$.

(2) In the **usual BIC approximation** the **"cost"** of each variable is 1 if included and 0 if not included in the model; we **generalize** this by bringing the actual costs in through the **prior**:

$$\boxed{(*)} \qquad f(\gamma_j) \propto \exp\left[-\frac{\gamma_j}{2}\left(\frac{c_j - c_0}{c_0}\right)\log n\right] \text{ for } j = 1, \ldots, p, \tag{7}$$

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

With $C_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j c_j$ as the **total cost** of model $\boldsymbol{\gamma}$, this yields our **generalized cost-adjusted** version of BIC,

$$-2\log PO_{k\ell} = -2\log\left[\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right] + \frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}\log n + O(1). \tag{8}$$

Approach (2) (continued)

MCMC implementation. We use a **two-step** method:

(1) First we use a **model search tool** to identify variables with **high marginal posterior inclusion probabilities** $f(\gamma_j|\boldsymbol{y})$, and we create a **reduced model space** consisting only of those variables whose marginal probabilities are above a **threshold value**.

(2) Then we use a **model search tool** in the **reduced model space** to estimate **posterior model probabilities** (and the corresponding odds).

We used both **RJMCMC** and **MCMC model composition** (MC^3) algorithm (Madigan-York, 1995) as model search tools; RJMCMC results below.

Results. The table below presents the **marginal posterior probabilities** of the variables that exceeded the threshold value of 0.30, in each of the **benefit-only** and **cost-benefit** analyses, together with their data collection costs (in minutes of abstraction time rather than US$), in the **Big World** of all 83 predictors.

In both the **benefit-only** and **cost-benefit** situations our methods reduced the initial list of $p = 83$ available candidates down to **13** predictors.

Results (continued)

Variable			Marginal Posterior Probabilities	
			Analysis	
Index	Name	Cost	Benefit-Only	Cost-Benefit
1	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

Note that the **most expensive** variables with high marginal posterior probabilities in the **benefit-only** analysis were **absent** from the set of promising variables in the **cost-benefit** analysis (e.g., Apache II Coma Score).

Results (continued)

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

Results (continued)

	Analysis		Percentage
	Benefit-Only	Cost-Benefit	Difference
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

- The minimum and median values of the posterior distribution of the **deviance** statistic for the benefit-only analysis were **lower** by a **relatively modest 5.3% and 5.1%** compared to the corresponding values of the cost-benefit analysis, but the **cost** of the best model in the cost-benefit analysis was almost **67% lower** than that for the benefit-only analysis; similarly, the dimensionality of the best model in the cost-benefit analysis was about **23% lower** than that for the benefit-only analysis.

These values indicate that the **loss of predictive accuracy** with the **cost-benefit analysis** is **small** compared to the **substantial gains** achieved in **cost** and **reduced model complexity**.

Utility Versus Cost-Adjusted BIC

Variable			Method		
			Utility	RJMCMC	
Index	Name	Cost (Minutes)	Good?	Good?	Posterior Probability
1	Systolic Blood Pressure Score (2-point scale)	0.5	**	**	0.99
2	Age	0.5	*	**	0.99
3	Blood Urea Nitrogen	1.5	**	**	1.00
4	APACHE II Coma Score (3-point scale)	2.5	**	**	1.00
5	Shortness of Breath Day 1 (yes, no)	1.0	**	**	0.99
6	Serum Albumin (3-point scale)	1.5	*	**	0.55
7	Respiratory Distress (yes, no)	1.0	*	**	0.92
8	Septic Complications (yes, no)	3.0			0.00
9	Prior Respiratory Failure (yes, no)	2.0			0.00
10	Recently Hospitalized (yes, no)	2.0			0.00
12	Initial Temperature	0.5	*	**	0.95
17	Chest X-ray Congestive Heart Failure Score (3-point scale)	2.5			0.00
18	Ambulatory Score (3-point scale)	2.5			0.00
48	Total APACHE II Score (36-point scale)	10.0			0.00

It's clear that the **utility** and **cost-adjusted BIC** approaches have reached **nearly identical conclusions** in the **Small World** of $p = 14$ predictors.

Utility Versus Cost-Adjusted BIC (continued)

With $p = 83$ the **agreement** between the two methods is also **strong** (although not as strong as with $p = 14$): using a **star system** for variable importance given in FND (2007a), **60** variables were **ignored** by both methods, **8** variables had **identical** star patterns, **3** variables were chosen as **important by both methods** but with different star patterns, **10** variables were marked as important by the utility approach and not by RJMCMC, and **2** variables were singled out by RJMCMC and not by utility: thus the two methods **substantially** agreed on the importance of **71 (86%) of the 83 variables**.

p	Method	Model	Cost	Median Deviance	LS_{CV}
14	RJMCMC	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_{12}$	9.0	1654	-0.329
		$X_1 + X_2 + X_3 + X_4 + X_5 + X_7 + X_{12}$	7.5	1676	-0.333
	Utility	$X_1 + X_3 + X_4 + X_5$	5.5	1726	-0.342
83	RJMCMC	$X_1 + X_2 + X_3 + X_5 + X_{12}$ $+ X_{46} + X_{49} + X_{51} + X_{70} + X_{78}$	7.5	1645	-0.327
	Utility	$X_1 + X_3 + X_4 + X_{12}$ $+ X_{46} + X_{49} + X_{57}$	6.5	1693	-0.336

To the extent that the two methods **differ**, the **utility** method favors models that **cost somewhat less** but also **predict somewhat less well**.

Utility Versus Cost-Adjusted BIC (continued)

The fact that the **two methods** may yield **somewhat different results** in **high-dimensional problems** does not mean that either is **wrong**; they are both **valid solutions** to **similar but not identical problems**.

Both methods lead to **noticeably better models** (in a **cost-benefit** sense) than frequentist or Bayesian **benefit-only** approaches, when — as is often the case — **cost** is an issue that must be included in the **problem formulation** to arrive at a **policy-relevant solution**.

Summary. In **comparing two or more models**, to say whether one is **better** than another I have to face the question: **better for what purpose**?

This makes **model specification** a **decision problem**: I need to either

(a) elicit a **utility structure** that's specific to the **goals** of the current study and **maximize expected utility** to find the best models, or

(b) (if (a) is too hard, e.g., because the problem has a **group decision** character) I can look for a **principled alternative** (like the **cost-adjusted BIC method** described here) that **approximates** the utility approach while **avoiding ambiguities in utility specification**.