Web supplement to *Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care*, by D. Fouskakis, I. Ntzoufras, and D. Draper

Imaginary data and power-prior motivation for the prior distribution in the main paper's equation (6)

After observing the design matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ for any model $\boldsymbol{\gamma}$, we consider a set of imaginary data $\boldsymbol{y}_i^* = (y_{i1}^* = 1,\ y_{i2}^* = 0), i = 1, \ldots, n$ that assigns probabilities 1/2 for all i and therefore supports the simplest (constant) model. We consider a prior that is generated using the likelihood of these imaginary data,

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y}^*) \propto \left\{ \prod_{i=1}^{n} p_i(\boldsymbol{\gamma})[1 - p_i(\boldsymbol{\gamma})] \right\}^{(2n)^{-1}}, \tag{37}$$

where $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*)$. Using the above prior, the posterior becomes

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y}) \propto \prod_{i=1}^{n} p_i(\boldsymbol{\gamma})^{y_i + \frac{1}{2n}} [1 - p_i(\boldsymbol{\gamma})]^{(1+\frac{1}{n}) - (y_i + \frac{1}{2n})}; \tag{38}$$

therefore this is equivalent to obtaining information from $\sum_{i=1}^{n}(1 + \frac{1}{n}) = (n+1)$ data points, instead of n data points when using a flat prior. Thus the proposed prior (37) introduces additional information to the posterior equivalent to adding one data point to the likelihood and therefore we support *a priori* the simplest model with a weight of one data point.

Using a Laplace approximation to (37) (see, e.g., Bernardo and Smith, 1994, p. 286), we obtain

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y}^*) \dot{\sim} N\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, 2n\,\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})^{-1}\right], \tag{39}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the maximum likelihood estimate if the imaginary data $\boldsymbol{y}_i^*$ were observed and $\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})$ is the observed information matrix given by

$$\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = \boldsymbol{X}_{\boldsymbol{\gamma}}^T \operatorname{diag}\left\{2\,\hat{p}_i^*(\boldsymbol{\gamma})[1 - \hat{p}_i^*(\boldsymbol{\gamma})]\right\} \boldsymbol{X}_{\boldsymbol{\gamma}}, \tag{40}$$

in which $\hat{p}_i^*(\boldsymbol{\gamma}) = \left[1 + \exp(-\boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\right]^{-1}$ is the fitted success probability for all i under model $\boldsymbol{\gamma}$ when observing data $\boldsymbol{y}^*$. Under the above imaginary data, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \boldsymbol{0}$ and $\hat{p}_i(\boldsymbol{\gamma}) = 1/2$ for all i, yielding $\mathcal{I}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = \frac{1}{2}\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}}\right)$ and therefore leading to the prior given by (6). 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 *a priori* support for more parsimonious models.

Details on RJMCMC and MC^3 implementation

The RJMCMC algorithm we employed can be summarized as follows:

1. For $j = 1, \ldots, p$, use RJMCMC to compare the current model $\boldsymbol{\gamma}$ with the proposed one $\boldsymbol{\gamma}'$ with components $\gamma_j' = 1 - \gamma_j$ and $\gamma_k' = \gamma_k$ for $k \neq j$ with probability one. The updating sequence of γ_j is randomly determined in each step.
2. For $j = 0, \ldots, p$, if $\gamma_j = 1$ then generate model parameters β_j from the corresponding posterior distribution $f(\beta_j|\boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}, \boldsymbol{y})$; otherwise set $\beta_j = 0$.

In our context the MC^3 algorithm may be summarized by the following steps:

1. For $j = 1, \ldots, p$, propose a move from the current model $\boldsymbol{\gamma}$ to a new one $\boldsymbol{\gamma}'$ with components $\gamma_j' = 1 - \gamma_j$ and $\gamma_k' = \gamma_k$ for $k \neq j$ with probability one. The updating sequence of γ_j is randomly determined in each step.

2. Accept the proposed model $\boldsymbol{\gamma}'$ with probability

$$\alpha = \min\left[1, \frac{f(\boldsymbol{\gamma}'|\boldsymbol{y})}{f(\boldsymbol{\gamma}|\boldsymbol{y})}\right] = \min\left(1, PO_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}\right). \tag{41}$$

Details on utility elicitation in Fouskakis and Draper (2008)

Since data on future patients are not available, Fouskakis and Draper (2008, hereafter FD) use a *cross-validation* approach (e.g., Gelfand *et al.*, 1992, Hadorn *et al.*, 1992) 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; FD take $\frac{n_V}{n} = \frac{1}{3}$).

In the approach taken by FD (and using the notation of that paper in this supplemental material), utility is quantified in monetary terms, so that the data collection utility is simply the negative of the total amount of money required to gather data on the specified predictor subset. 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

$$U_D(I) = -n_V \sum_{j=1}^{p} c_j I_j, \tag{42}$$

where c_j is the marginal cost per patient of data abstraction for variable j.

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. 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. FD 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 input-output 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; see, e.g., Kahn, Rogers *et al.*, 1990).

In the first step of their approach, given a particular predictor subset I, FD fit a logistic regression model to the modeling subsample M and apply this model to the 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

$$\begin{array}{c} (Y_i \,|\, p_i^I) \overset{\text{indep}}{\sim} \text{Bernoulli}(p_i^I), \\ \log(\frac{p_i^I}{1-p_i^I}) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}. \end{array} \tag{43}$$

FD 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

$$\hat{p}_i^I = \left[1 + \exp\left(-\sum_{j=0}^{k} \hat{\beta}_j X_{ij}\right)\right]^{-1}, \tag{44}$$

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 their approach FD 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. FD 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 Table 8. The left-hand part of this table records the rewards and penalties in US$. The predictive utility of model I is then

$$U_P(I) = \sum_{l=1}^{2} \sum_{m=1}^{2} C_{lm} \, n_{lm}. \tag{45}$$

Table 8: *Cross-tabulation of actual versus predicted death status. The left-hand table records the monetary rewards and penalties for correct and incorrect predictions; the right-hand table summarizes the frequencies in the* 2×2 *tabulation.*

		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}

To elicit the utility values C_{lm} FD reason as follows. (1) Clearly C_{11} (the reward for correctly predicting death at 30 days) and C_{22} (the reward for correctly predicting living at 30 days) should be positive, and C_{12} (the penalty for a false prediction of living) and C_{21} (the penalty for a false prediction of death) should be negative. (2) Since it is easier to correctly predict that a person lives than dies with these data (the overall pneumonia 30–day death rate in the RAND sample was 16%, so a prediction that every patient lives would be right about 84% of the time), it is natural to specify that $C_{11} > C_{22}$. (3) Since it is arguably worse to label a "bad" hospital as "good" than the other way around, one should take $|C_{12}| > |C_{21}|$, and furthermore it is natural that the magnitudes of the penalties should exceed those of the rewards. (4) FD completed the utility specification by eliciting information from health experts in the U.S. and U.K, first to anchor C_{21} to the cost of subjecting a "good" hospital to an unnecessary process audit and then to obtain ratios relating the other C_{lm} to C_{21}.

Since the utility structure used in FD is based on the idea that hospitals have to be treated in an all-or-nothing way in acting on the basis of their apparent quality, the approach taken was (i) to attempt to quantify the monetary loss L of incorrectly subjecting a "good" hospital to a detailed but unnecessary process audit and then (ii) to translate this from the hospital to the patient level. A rough correspondence may be made between the left-hand part of Table 8 at the patient level and a hospital-level table with rows representing truth ("bad" in row 1, "good" in row 2) and columns representing the decision taken ("process audit" in column 1, "no process audit" in column 2). Unnecessary process audits then correspond to cell $(2, 1)$ in these tables (hospitals where a process audit is not needed will typically have an excess of patients who are predicted to die but actually live). Discussions with health experts in the U.S. and U.K. suggested that detailed process audits cost on the order of $L = \$5,000$ per hospital (in late 1980s U.S. dollars), and RAND data indicated that the mean number of pneumonia patients per hospital per year in the U.S. at the time of the RAND quality of care study was 71.8. This fixed C_{21} at approximately $\frac{-\$5{,}000}{71.8} = -\$69.6$. FD's health experts judged that C_{12} should be the largest in absolute value of the C_{lm}, and averaging across the expert opinions, expressed as orders of magnitude base 2, the elicitation results were $\left|\frac{C_{12}}{C_{21}}\right| = 2$, $\left|\frac{C_{11}}{C_{21}}\right| = \frac{1}{2}$, and $\left|\frac{C_{22}}{C_{21}}\right| = \frac{1}{8}$, finally yielding $(C_{11}, C_{12}, C_{21}, C_{22}) = \$(34.8, -139.2, -69.6, 8.7)$. The results in FD and this paper 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

$$E\left[U(I)\right] = E\left[U_D(I) + U_P(I)\right], \tag{46}$$

where this expectation is over all possible cross-validation splits of the data. The number of possible cross-validation splits is far too large to evaluate the expectation in (46) directly; in practice FD therefore use Monte Carlo methods to evaluate it, averaging over N random modeling and validation splits.

References in the web supplement

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.

Hadorn D, Draper D, Rogers W, Keeler E, Brook R (1992). Cross-validation performance of patient mortality prediction models. *Statistics in Medicine*, **11**, 475–489.