

Empirical Bayes vs. Fully Bayes Variable Selection

Wen CUI and Edward I. GEORGE *

December 2004, Revised December 2006

Abstract

For the problem of variable selection for the normal linear model, fixed penalty selection criteria such as AIC, C_p , BIC and RIC correspond to the posterior modes of a hierarchical Bayes model for various fixed hyperparameter settings. Adaptive selection criteria obtained by empirical Bayes estimation of the hyperparameters have been shown by George and Foster (2000) to improve on these fixed selection criteria. In this paper, we study the potential of alternative fully Bayes methods, which instead margin out the hyperparameters with respect to prior distributions. Several structured prior formulations are considered for which fully Bayes selection and estimation methods are obtained. Analytical and simulation comparisons with empirical Bayes counterparts are studied.

**Wen Cui is Assistant Professor, Texas State University, San Marcos, Texas 78666 jcu@txstate.edu. Edward I. George is Professor, The Wharton School, 3730 Walnut Street 400 JMHH, Philadelphia, PA 19104-6340, edgeorge@wharton.upenn.edu. We would like to thank the Editor, two anonymous referees and Merlise Clyde for helpful comments and suggestions. This work was supported by a number of NSF grants, where DMS-0605102 is the most recent.*

1 Introduction

We consider the variable selection problem in the context of normal linear regression. Suppose the relationship between a dependent variable, Y , and potential explanatory variables, X_1, \dots, X_p , can be described by a normal linear model,

$$Y = X\beta + \epsilon, \quad (1)$$

in which Y is $n \times 1$, X is $n \times p$, and $\epsilon \sim N_n(0, \sigma^2 I)$. The variable selection problem arises when there is uncertainty about which, if any, of the explanatory variables should be dropped from the model. Letting $\gamma = 1, \dots, 2^p$ index the subsets of X_1, \dots, X_p , and letting X_γ be the $n \times q_\gamma$ design matrix corresponding to the γ th subset, this corresponds to uncertainty about which is the appropriate subset model

$$Y = X_\gamma \beta_\gamma + \epsilon. \quad (2)$$

A common strategy under such variable selection uncertainty has been to select the γ th model which maximizes a penalized regression sum of squares criterion of the form

$$SS_\gamma / \hat{\sigma}^2 - F q_\gamma. \quad (3)$$

Here $SS_\gamma = \hat{\beta}_\gamma^{LS} X_\gamma' X_\gamma \hat{\beta}_\gamma^{LS}$ is the regression sum of squares of the γ th model, where $\hat{\beta}_\gamma^{LS}$ is the conditional least squares estimate of β_γ , $\hat{\sigma}^2$ is an estimate of σ^2 (such as the classical unbiased estimate based on the full model) and F is a fixed penalty for adding a variable. Such criteria include AIC (Akaike 1973) and C_p (Mallows 1973) when $F = 2$; BIC (Schwarz 1978) when $F = \log n$; and RIC (Foster and George 1994) when $F = 2 \log p$.

More recently, a wide variety of semiautomatic and objective Bayesian approaches to the variable selection problem have appeared in the literature, for example, see Berger and Pericchi (2001), Chipman, George and McCulloch (2001), Clyde and George (2004), Casella and Moreno (2006) and the references therein. Of particular interest for this paper, are the Empirical Bayes (EB) criteria proposed by George and Foster (2000) which were further developed and extended by Clyde and George (2000) and Johnstone and Silverman (2004).

George and Foster's EB criteria also entail maximization of a penalized sum of squares such as (3) but with F replaced by adaptive dimensionality penalties that are obtained via a Bayesian calibration to the data. The model for this

calibration uses the hierarchical mixture prior on β_γ and γ ,

$$p(\beta_\gamma, \gamma \mid c, \omega) = p(\beta_\gamma \mid \gamma, c) p(\gamma \mid \omega) \quad (4)$$

where

$$p(\beta_\gamma \mid \gamma, c) = N_{q_\gamma}(0, c (X_\gamma' X_\gamma)^{-1} \sigma^2), \quad c > 0, \quad (5)$$

and

$$p(\gamma \mid \omega) = \omega^{q_\gamma} (1 - \omega)^{p - q_\gamma}, \quad \omega \in [0, 1]. \quad (6)$$

This prior is determined by two hyperparameters c and ω , where c controls the average size of the β_γ , and ω controls the average number of nonzero coefficients in the model. Note that increasing c and/or ω serves to increase the overall “strength of the signal”.

The connection between the induced hierarchical Bayes model and penalty criteria of the form (3) is revealed by the posterior for γ , namely

$$p(\gamma \mid Y, c, w) \propto \exp \left\{ \frac{c}{2(1+c)} [SS_\gamma / \sigma^2 - F(c, w) q_\gamma] \right\}, \quad (7)$$

where

$$F(c, w) = \frac{1+c}{c} \left\{ 2 \log \frac{1-w}{w} + \log(1+c) \right\}. \quad (8)$$

For fixed Y , c and ω , $p(\gamma \mid Y, c, w)$ is increasing in

$$SS_\gamma / \sigma^2 - F(c, w) q_\gamma, \quad (9)$$

which is precisely (3) with $F = F(c, w)$. Thus, selecting the highest posterior model under this prior is equivalent to selecting the model maximizing (3) with $F = F(c, w)$. By suitable choices of c and ω , the posterior mode can be calibrated to correspond to traditional fixed penalty criteria such as AIC/Cp, BIC or RIC. respectively.

Rather than using fixed prespecified values for c and ω , George and Foster (2000) considered estimating them from the data via empirical Bayes. For this purpose, they proposed two approaches, marginal maximum likelihood (MML) and conditional maximum likelihood (CML). MML entails finding \hat{c} and $\hat{\omega}$ that maximize the overall marginal likelihood

$$\begin{aligned} L(c, w \mid Y) &\propto \sum_{\gamma} p(\gamma \mid w) p(Y \mid \gamma, c) \\ &\propto \sum_{\gamma} \omega^{q_\gamma} (1 - \omega)^{p - q_\gamma} (1 + c)^{-q_\gamma/2} \exp \left\{ \frac{c SS_\gamma}{2\sigma^2(1+c)} \right\}, \quad (10) \end{aligned}$$

and inserting them into (9) to obtain

$$C_{\text{MML}} = SS_\gamma / \sigma^2 - F(\hat{c}, \hat{w}) q_\gamma. \quad (11)$$

Note that the penalty $F(\hat{c}, \hat{w})$ adapts to the data through the estimates of c and w . George and Foster (2000) showed via simulations that, as opposed to fixed penalty criteria, the performance of C_{MML} is nearly as good as the best possible fixed penalty criterion over a broad range of model specifications.

A drawback of C_{MML} is that it can be computationally overwhelming especially when X is nonorthogonal because maximizing (10) involves averaging γ over the whole model space. To mitigate this difficulty George and Foster (2000) also proposed C_{CML} , an easily computable alternative. C_{CML} entails choosing the model γ for which the conditional likelihood

$$\begin{aligned} L^*(c, w, \gamma | Y) &\propto p(\gamma | w) p(Y | \gamma, c) \\ &\propto w^{q_\gamma} (1 - w)^{p - q_\gamma} (1 + c)^{-q_\gamma/2} \exp \left\{ \frac{c SS_\gamma}{2\sigma^2(1 + c)} \right\} \end{aligned} \quad (12)$$

is maximized over c , w and γ . We further discuss the form of C_{CML} in Section 2.3. Although its performance was not quite as good as that of C_{MML} , George and Foster (2000) showed that C_{CML} offered similar adaptive improvements over fixed penalty criteria.

The main thrust of this paper is to propose and explore the potential of some Fully Bayes (FB) alternatives to these EB criteria. As opposed to EB estimation of c and w , FB approaches put hyperpriors on c and w and then integrate them out to obtain the marginal posterior, $\pi(\gamma | Y)$ over the model space. As with C_{MML} and C_{CML} , the FB posterior mode can then be used for model selection.

For particular conjugate hyperpriors, we obtain nearly closed, easily computable forms for these posteriors which we then use for analytical and performance comparisons with C_{MML} and C_{CML} . Because in many statistical decision problems, the admissible estimators are either Bayes or limits of Bayes procedures (Berger 1985), one might anticipate that such FB procedures would improve over C_{MML} and C_{CML} which are neither Bayes nor limits of Bayes procedures. Surprisingly, it appears that our FB procedures are inferior to C_{MML} and essentially comparable to C_{CML} .

Throughout this paper we condition on σ and treat it as if it were known. We do this because eliminating this source of uncertainty allows us to more

clearly compare the EB and FB procedures both analytically and in terms of their simulated performance, which is our main thrust. In practice, of course, it is necessary and important to treat unknown σ . As we discuss in Section 4, this can be done for the EB and FB procedures by using plug-in estimates of σ . However, a further potential advantage of the FB approach is that it also straightforwardly allows for a Bayesian treatment of unknown σ . Such practical fully Bayesian approaches are considered and studied in Liang, Paulo, Molina, Clyde and Berger (2006).

The structure of this paper is as follows. In Section 2, various priors for c and ω are discussed and considered; under the priors, the FB selection criterion is derived and compared with the corresponding EB criteria; FB conditional posterior mean estimates of β are obtained for inference after selection. In Section 3, the performance of the various procedures are compared via simulations. In Section 4, we conclude with a discussion of our findings.

2 Fully Bayes Selection Criteria

The formulation of an FB selection procedure is straightforward in principle: hyperpriors are chosen for c and w and then these two hyperparameters are integrated out. FB selection then simply entails selecting the model with highest posterior probability. We begin with a discussion of the choice of the hyperpriors.

2.1 Hyperpriors on ω and c

In seeking hyperpriors for c and w , several characteristics are especially desirable. First of all, because we are ultimately considering a Bayesian model selection problem, it is crucial to use proper hyperpriors. This avoids the problem of arbitrary normalizing constants that would render Bayes factors meaningless. Secondly, because of the typically large number of models in variable selection problems, it is advantageous to have hyperpriors leading to computationally tractable posteriors. In particular, hyperpriors that lead to closed form posterior expressions are ideal. Finally, because very little, if any, subjective prior information is available in model selection settings such as ours, it is especially appealing to have objective prior formulations that can serve as default settings for automatic use. To satisfy all three of these desiderata, we turn to conjugate

hyperpriors. These are proper, they yield closed form posteriors, and allow for some reasonable default settings.

We begin with the simple conjugate prior family for w , namely the Beta(w_a, w_b) distributions, which yield

$$\pi_{w_a, w_b}(\gamma) = \frac{\Gamma(q_\gamma + w_a) \Gamma(p - q_\gamma + w_b)}{\Gamma(p + w_a + w_b)} \frac{\Gamma(w_a + w_b)}{\Gamma(w_a) \Gamma(w_b)}. \quad (13)$$

With mean $w_a/(w_a + w_b)$ and variance decreasing in w_a and w_b , the hyperparameters w_a and w_b can be chosen to reflect a preference towards particular models. For example, $w_a/(w_a + w_b)$ small would reflect a preference for parsimonious models. For a default choice, three natural contenders are $w_a = w_b = 1$ which yields the uniform hyperprior on w , $w_a = w_b = -\frac{1}{2}$ which yields the Jeffreys prior, and $w_a = w_b = -1$ which yields the Haldane prior. For a discussion of the relative virtues of these three priors, see Geisser (1984), who preferred $w_a = w_b = 1$, and the references therein.

As the default of π_{w_a, w_b} , we consider $w_a = w_b = 1$, which yields

$$\pi_{1,1}(\gamma) = \frac{1}{p+1} \left(\begin{matrix} p \\ q_\gamma \end{matrix} \right)^{-1}, \quad (14)$$

a special case of a form proposed by George and McCulloch (1993). Note that under (14), the prior on model size

$$\pi_{1,1}(q_\gamma) = \frac{1}{p+1} \quad (15)$$

is uniform. This stands in stark contrast to the popular uniform prior on γ ,

$$\pi_U(\gamma) \equiv \frac{1}{2^p} \quad (16)$$

which induces the prior on model size

$$\pi_U(q_\gamma) = \frac{1}{2^p} \left(\begin{matrix} p \\ q_\gamma \end{matrix} \right). \quad (17)$$

Compared to π_U , $\pi_{1,1}$ places much more weight on the sparse and the saturated models. Note that π_U is the limiting distribution of π_{w_a, w_b} as $w_a = w_b \rightarrow \infty$.

Turning to c , we note that $(1+c)$ serves as a scale parameter in the marginal distribution of $f(y|c, \gamma)$. Thus, a natural choice is the incomplete Inverse Gamma (α, b) conjugate form for $(1+c)$ for $c \in (0, \infty)$. This yields the hyperprior for c

$$\pi_{\alpha, b}(c) = M(1+c)^{-(1+\alpha)} \exp \left\{ -\frac{b}{1+c} \right\}, \quad (18)$$

where $M = b^\alpha \left(\int_0^b t^{\alpha-1} e^{-t} dt \right)^{-1}$ and $c \in (0, \infty)$. As will be seen in Section 2.2, this prior leads to nearly closed form posterior expressions involving only a single one dimension integral.

The prior in (18) is controlled by two hyperparameters, α and b that control its shape and scale. As α and b are increased the prior becomes less flat and more concentrated near 0. Thus, in the spirit of stable estimation, a natural default choice would entail using small values for α and b . For this purpose, we recommend setting $b = 0$ and using

$$\pi_\alpha(c) = \alpha(1+c)^{-(1+\alpha)} \text{ for } c \in (0, \infty) \quad (19)$$

with $\alpha = 1$ as the default prior on c . Although even smaller choices for α might be considered, extensive simulation studies in Cui (2002) suggest that the performance of the FB procedures, which we discuss below, is robust to small changes around $\alpha = 1$. It might be noted that $\alpha = 0$ and $b = 0$ yields an improper prior which is proportional to the Jeffrey's prior conditional on γ . Because it is not proper, such a prior does not appear to be useful for model selection.

Priors of the form (19) were also recently and independently proposed for the variable selection problem by Liang et. al. (2006). Referring to them as hyper-g priors, they studied these for the important practical case of unknown variance in conjunction with $\pi(\sigma^2) \propto 1/\sigma^2$. Priors of the form (19) were also proposed by Strawderman (1971) for the somewhat different context of minimax estimation of a multivariate normal mean with identity covariance.

As an alternative to priors of the form (18), one might instead consider Inverse Gamma hyperpriors on c rather than on $(1+c)$. Such a conjugate choice is implicit in the work of Zellner and Siow (1980, 1984). Motivated by Jeffreys (1967), Zellner and Siow proposed testing $H_1 : \beta_\gamma = 0$ versus $H_2 : \beta_\gamma \neq 0$ using multivariate Cauchy priors with covariance $X'_\gamma X_\gamma/n$ for $\pi(\beta_\gamma | \sigma)$ and $\pi(\sigma) \propto 1/\sigma^2$. Such priors would be obtained here by putting an inverse Gamma hyperprior $IG(\frac{1}{2}, \frac{n}{2})$ directly on c . Unfortunately, posteriors under this prior are computationally more difficult to approximate than posteriors under (19). Zellner-Siow priors for variable selection with unknown σ were also extensively studied by Liang et. al. (2006).

2.2 Fully Bayes Posteriors

Under the $Beta(w_a, w_b)$ hyperprior for w and the incomplete gamma conjugate form (18) for c , the Normal-Bernoulli setup (5)–(6), and the linear model (1), the model posterior is straightforwardly obtained as

$$\pi(\gamma | Y) = K \cdot M \cdot \exp \left\{ \frac{SS_\gamma}{2\sigma^2} \right\} \left(b + \frac{SS_\gamma}{2\sigma^2} \right)^{-\frac{q_\gamma}{2} - \alpha} G_{\alpha, b}(SS_\gamma, q_\gamma) \pi_{w_a, w_b}(\gamma) \quad (20)$$

for $q_\gamma \neq 0$ and

$$\pi(\gamma | Y) = K \cdot \pi_{w_a, w_b}(\gamma) \quad (21)$$

for $q_\gamma = 0$, where

$$K = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{Y'Y}{2\sigma^2} \right\} / m(Y), \quad (22)$$

$$G_{\alpha, b}(SS_\gamma, q_\gamma) = \int_0^{\frac{SS_\gamma}{2\sigma^2} + b} t^{\frac{q_\gamma}{2} + \alpha - 1} e^{-t} dt, \quad (23)$$

$m(Y)$ is the overall marginal density function of Y , and M is the norming constant in (18).

For our default prior choices $\pi_{1,1}(\gamma)$ in (14) and $\pi_1(c)$ in (19), the model posterior reduces to

$$\pi(\gamma | Y) = K \cdot \exp \left\{ \frac{SS_\gamma}{2\sigma^2} \right\} \left(\frac{SS_\gamma}{2\sigma^2} \right)^{-\frac{q_\gamma}{2} - 1} G(SS_\gamma, q_\gamma) \pi_{1,1}(\gamma) \quad (24)$$

for $q_\gamma \neq 0$ and

$$\pi(\gamma | Y) = \frac{K}{p+1} \quad (25)$$

for $q_\gamma = 0$, where

$$G(SS_\gamma, q_\gamma) = \int_0^{\frac{SS_\gamma}{2\sigma^2}} t^{\frac{q_\gamma}{2}} e^{-t} dt. \quad (26)$$

Model selection under any one of these posteriors can then simply proceed by choosing the model γ that maximizes $\pi(\gamma | Y)$. For comparisons with empirical Bayes criteria, we focus on the default Fully Bayes criterion obtained by maximizing (24) and (25).

2.3 Comparison of C_{FB} and C_{CML}

It is interesting to compare the default Fully Bayes criterion that maximizes (24) with the Empirical Bayes criterion that maximizes (12). To do this, we consider the penalized sum of squares representation of the Empirical Bayes criterion

$$C_{\text{CML}} = \begin{cases} SS_{\gamma}/\sigma^2 - B(SS_{\gamma}, q_{\gamma}) - R(q_{\gamma}) & \text{if } q_{\gamma} \neq 0 \\ 0 & \text{if } q_{\gamma} = 0 \end{cases} \quad (27)$$

where letting $I_{\gamma} = 1$ if $SS_{\gamma}/\sigma^2 q_{\gamma} > 1$ and $I_{\gamma} = 0$ otherwise,

$$B(SS_{\gamma}, q_{\gamma}) = I_{\gamma} q_{\gamma} \left\{ 1 + \log \left(\frac{SS_{\gamma}}{\sigma^2 q_{\gamma}} \right) \right\} + (1 - I_{\gamma}) \frac{SS_{\gamma}}{\sigma^2} \quad (28)$$

and

$$R(q_{\gamma}) = -2 \{ (p - q_{\gamma}) \log(p - q_{\gamma}) + q_{\gamma} \log q_{\gamma} \}. \quad (29)$$

As shown by George and Foster (2000), the component B is a consequence from estimating c and acts like BIC, whereas the component R is a consequence from estimating ω and acts like RIC. (The expression for C_{CML} above corrects a minor error in the expression for C_{CML} in George and Foster (2000). However, the error is relatively unimportant as it occurs only when $I_{\gamma} = 0$, an event of low probability).

By maximizing $2 \log \pi(\gamma | Y)$ with irrelevant constants removed, we can express the default Fully Bayes criterion as an analogous penalized sum of squares criterion, denoted C_{FB} ,

$$C_{\text{FB}} = \begin{cases} SS_{\gamma}/\sigma^2 - B^*(SS_{\gamma}, q_{\gamma}) - R^*(q_{\gamma}) & \text{if } q_{\gamma} \neq 0 \\ 0 & \text{if } q_{\gamma} = 0 \end{cases} \quad (30)$$

where

$$B^*(SS_{\gamma}, q_{\gamma}) = (q_{\gamma} + 2) \log \frac{SS_{\gamma}}{2\sigma^2} - 2 \log G(SS_{\gamma}, q_{\gamma}) \quad (31)$$

and

$$R^*(q_{\gamma}) = -2 \{ \log(p - q_{\gamma})! + \log q_{\gamma}! - \log(p + 1)! \}. \quad (32)$$

Analogous to the C_{CML} penalties B and R , here B^* and R^* are the penalties due to marginalizing over c and w , respectively.

It is interesting to compare the respective penalties of C_{CML} and C_{FB} when the model dimension goes from $q_{\gamma} - 1$ to q_{γ} with a negligible change in SS_{γ} ,

corresponding to the addition of an unimportant variable. Assuming no change in SS_γ and that $SS_\gamma/\sigma^2 q_\gamma > 1$, the change in B is obtained as

$$\Delta B(SS_\gamma, q_\gamma) = \left(1 + \log \frac{SS_\gamma}{\sigma^2 q_\gamma}\right) - (q_\gamma - 1) \log \frac{q_\gamma}{q_\gamma - 1}. \quad (33)$$

Under the same assumptions, the change in B^* is

$$\Delta B^*(SS_\gamma, q_\gamma) = \log \frac{SS_\gamma}{2\sigma^2} - 2 \log \frac{G(SS_\gamma, q_\gamma)}{G(SS_\gamma, q_\gamma - 1)} \quad (34)$$

$$\approx \log \frac{SS_\gamma}{2\sigma^2} - 2 \log \frac{\frac{SS_\gamma}{2\sigma^2} \left(\frac{q_\gamma}{2}\right)^{\frac{q_\gamma}{2}} e^{-\frac{q_\gamma}{2}}}{\frac{SS_\gamma}{2\sigma^2} \left(\frac{q_\gamma - 1}{2}\right)^{\frac{q_\gamma - 1}{2}} e^{-\frac{q_\gamma - 1}{2}}} \quad (35)$$

$$= \left(1 + \log \frac{SS_\gamma}{\sigma^2 q_\gamma}\right) - (q_\gamma - 1) \log \frac{q_\gamma}{q_\gamma - 1}, \quad (36)$$

where we have used the upper bound approximation

$$G(SS_\gamma, x) = \int_0^{\frac{SS_\gamma}{2\sigma^2}} t^{\frac{x}{2}} e^{-t} dt \approx \frac{SS_\gamma}{2\sigma^2} \left(\frac{x}{2}\right)^{\frac{x}{2}} e^{-\frac{x}{2}}, \quad (37)$$

($t^x e^{-t}$ is maximized at $t = x$). Note that when $SS_\gamma/\sigma^2 q_\gamma > 1$, $q_\gamma/2$ is within the range of integration, thereby improving the quality of this approximation.

Turning to the relative changes in R and R^* , George and Foster (2000) use the approximation

$$1 + \log q \approx q \log q - (q - 1) \log(q - 1) \quad (38)$$

to show that

$$\Delta R(q_\gamma) \approx 2 \log \frac{p - q_\gamma + 1}{q_\gamma}. \quad (39)$$

For the C_{FB} criterion, this approximation turns out to be exact

$$\Delta R^*(q_\gamma) = 2 \log \frac{p - q_\gamma + 1}{q_\gamma}. \quad (40)$$

Thus we see that, at least for small changes in SS_γ , the penalties imposed by C_{CML} and C_{FB} are very similar.

2.4 Estimation of β_γ After Selection

When C_{FB} is used to select a model γ , it will usually also be of interest to estimate the corresponding vector of coefficients β_γ . For this purpose in the fully Bayes framework, it is most natural to use the conditional posterior mean

of β_γ given the selected γ , namely $E(\beta_\gamma | Y, \gamma)$. Under our default priors, an easily computable expression for this is obtained as

$$\hat{\beta}_\gamma^{FB} = E(\beta_\gamma | \hat{\beta}_\gamma^{LS}, \gamma) = \left(1 - \frac{2\sigma^2}{SS_\gamma} \frac{G(SS_\gamma, q_\gamma + 2)}{G(SS_\gamma, q_\gamma)}\right) \hat{\beta}_\gamma^{LS}, \quad (41)$$

where $\hat{\beta}_\gamma^{LS}$ is the conditional least squares estimate of β_γ . This follows using $E(\beta_\gamma | Y, \gamma) = E(E(\beta_\gamma | \hat{\beta}_\gamma^{LS}, \gamma, c))$ and the fact that $E(\beta_\gamma | \hat{\beta}_\gamma^{LS}, \gamma, c) = \frac{c}{1+c} \hat{\beta}_\gamma^{LS}$ under the normal prior (5).

In contrast, in the empirical Bayes framework, one further conditions on the estimate \hat{c} . For C_{MML} , this yields

$$\hat{\beta}_\gamma^{MML} = E(\beta_\gamma | Y, \gamma, \hat{c}) = \frac{\hat{c}}{1 + \hat{c}} \hat{\beta}_\gamma^{LS}. \quad (42)$$

where \hat{c} the maximum marginal likelihood estimate of c . Although this can be numerically computed, no closed form representation for $\hat{\beta}_\gamma^{MML}$ is available. However, for C_{CML} , the EB posterior mean is obtained in closed form as

$$\hat{\beta}_\gamma^{CML} = \left(1 - \frac{\sigma^2 q_\gamma}{SS_\gamma}\right)_+ \hat{\beta}_\gamma^{LS}, \quad (43)$$

where $(a)_+$ denotes the positive part of a . The approximation

$$\hat{\beta}_\gamma^{FB} \approx \left(1 - \frac{\sigma^2(q_\gamma + 2)}{SS_\gamma}\right)_+ \hat{\beta}_\gamma^{LS}, \quad (44)$$

which is obtained from (41) using (37) and then (38), reveals that $\hat{\beta}_\gamma^{CML}$ and $\hat{\beta}_\gamma^{FB}$ are very similar. Note that all three of these posterior mean estimators shrink $\hat{\beta}_\gamma^{LS}$ towards zero.

3 Simulation Evaluations

We now turn to simulation comparisons of our default fully Bayes procedures C_{FB} with the empirical Bayes procedures C_{MML} and C_{CML} . To shed light on the effect of the model space prior $\pi_{1,1}(\gamma)$ in (14), we have included the default fully Bayes procedure which, like C_{FB} , uses $\pi_1(c)$ in (19) but replaces $\pi_{1,1}(\gamma)$ by the uniform $\pi_U(\gamma)$ in (16). The resulting criterion to be maximized, which we denote C_{FBU} , differs from C_{FB} in (30) only in that $R^*(q_\gamma)$ in (32) is replaced by a constant. Finally, for added perspective we also included AIC and BIC, the traditional penalizes sum-of-squares criteria with fixed penalties $F = 2$ and $F = \log n$, respectively.

From a decision theoretic point of view, the appropriate loss for such selection rules is the 0-1 loss function which is 0 if and only if the correct model is selected. However, using such a loss function for simulation is problematic because getting the model exactly right is a very small probability event when so many models are being compared. Furthermore, such a loss function ignores the extent to which the selected model differs from the correct model; getting most of the variables correct is not rewarded at all.

Thus, as in George and Foster (2000), we find it more informative to summarize the performance of the various selection rules by the predictive error loss

$$L\{\beta, \hat{\beta}_\gamma\} \equiv \{X\hat{\beta}_\gamma - X\beta\}'\{X\hat{\beta}_\gamma - X\beta\} \quad (45)$$

where $\hat{\beta}_\gamma$ is an estimator of β_γ conditionally on the selected γ . For the purpose of comparing the selection criteria C_{MML} , C_{CML} , C_{FB} , C_{FBU} , AIC and BIC, we use $\hat{\beta}_\gamma = \hat{\beta}_\gamma^{LS}$. For the purpose of evaluating and comparing the estimation-after-selection posterior mean rules, $\hat{\beta}_\gamma^{MML}$, $\hat{\beta}_\gamma^{CML}$, $\hat{\beta}_\gamma^{FB}$ and $\hat{\beta}_\gamma^{FBU}$, we substitute each of these for $\hat{\beta}_\gamma$ in (45). Note that $\hat{\beta}_\gamma^{FBU}$ is obtained just as $\hat{\beta}_\gamma^{FB}$ in (41) except that γ is instead selected using C_{FBU} .

For simplicity, we assume that $\sigma^2 = 1$ is known, and focus exclusively on the orthogonal case where $X = I$, a setting that arises naturally in nonparametric regression. In this case, the regression model (2) is of the simple form $Y = \beta + \epsilon$. Following George and Foster (2000), we generated data as follows. For $n = p = 1000$ and fixed values of q and c , we generated the first q components of β as independent $N(0, c)$ realizations, set $\beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$, independently generated $\epsilon \sim N_p(0, I_p)$, and then calculated $Y = \beta + \epsilon$. This procedure was replicated 1000 times and the average loss (45) was calculated for each procedure. These average losses were obtained for each value of $q = 0, 10, 25, 50, 100, 300, 500, 700, 900, 1000$ when $c = 5, 25$. Note that $c = 5$ and $c = 25$ respectively correspond to weak and strong signal-to-noise ratios.

We remark that for each fixed q_γ , the posterior $\pi(\gamma | Y)$ in (24), is monotonically increasing in SS_γ . This property allows us to easily identify the highest posterior models when X is orthogonal by simply using greedy forward selection of Y coordinates to obtain the sequence of maximal SS_γ values for $q = 0, \dots, 1000$. If X is nonorthogonal and the problem is large, one must resort to methods that heuristically restrict attention to a subset of the model space, and then apply the selection criteria to the subset.

Before proceeding, we should point out a feature that C_{FB} shares with C_{CML} , namely a tendency towards bimodality over the γ space. This is illustrated by the six plots in Figure 1 which display the maximum of the log posterior (24) for each model size q over 50 simulations of six of our model setups. Note that for moderate size models, true model size tends to fall between location of the two nodes. These plots were typical of what we saw in most of the simulations. In sharp contrast, Figure 2 reveals no bimodality whatsoever in the log posteriors corresponding to C_{FBU} for the identical data. This shows quite clearly that the bimodality of C_{FB} is entirely due to effect of the prior $\pi_{1,1}(\gamma)$ which, as was pointed out in Section 2.1, puts much larger weight on the sparse (small q_γ) and saturated (large q_γ) models. Because C_{CML} is implicitly using approximately this prior, this evidently also explains the posterior bimodality of C_{CML} , rather than the speculative bimodality explanation given by George and Foster (2000). To mitigate the bimodality difficulty in practice, we recommend selecting the more parsimonious of the two models when such bimodality is present. This modification, also used by George and Foster (2000) for C_{CML} , helped to improve the performance of C_{FB} for smaller q models.

Table 1 presents the average losses over 1000 simulations of the setup described above, for the EB selection rules C_{MML} , C_{CML} , the FB selection rules C_{FB} , C_{FBU} , the fixed penalty selection rules AIC and BIC, and the estimation-after-selection posterior mean rules $\hat{\beta}_\gamma^{\text{MML}}$, $\hat{\beta}_\gamma^{\text{CML}}$, $\hat{\beta}_\gamma^{\text{FB}}$ and $\hat{\beta}_\gamma^{\text{FBU}}$. For visual comparisons, these selection rule losses are plotted in Figure 3 and the posterior mean rule losses are plotted in Figure 4.

The main focus of our investigation is the extent to which C_{FB} compares with C_{MML} and C_{CML} . We can see immediately from Table 1 and Figure 3 that when $c = 5$, corresponding to a low signal-to-noise ratio, C_{MML} performed markedly better than both C_{FB} and C_{CML} except for very small values of q . When $c = 25$, C_{MML} was still best, although by a lesser extent, again except for very small values of q . It is also clear that, just as the analytical comparisons in Section 2.3 suggested, C_{FB} and C_{CML} are essentially equivalent, although C_{FB} seems to perform very slightly better than C_{CML} .

Turning to C_{FBU} , Table 1 and Figure 3 reveal that replacing $\pi_{1,1}(\gamma)$ by the uniform $\pi_U(\gamma)$ has a substantial effect. C_{FB} performed substantially better than C_{FBU} for smaller q , and for larger q when $c = 25$. Compared to AIC and BIC, which we included simply to add perspective, C_{MML} , C_{CML} and C_{FB} are all clearly

Average losses when $c = 5$

q	0	10	25	50	100	300	500	700	900	1000
C_{MML}	3.7	36.7	79.2	144.1	259.5	625.3	878.9	998.0	1000.3	1001.8
C_{CML}	0.3	35.1	81.1	149.9	273.6	682.2	990.3	1210.6	1301.9	1170.7
C_{FB}	0.3	35.2	81.0	149.3	273.2	681.4	989.7	1209.9	1301.2	1169.6
C_{FBU}	633.0	625.5	618.4	614.0	618.4	715.2	864.0	1024.9	1188.7	1274.2
AIC	572.6	577.0	586.4	603.7	636.5	755.3	879.1	1002.1	1125.5	1188.1
BIC	75.7	93.3	120.6	169.2	261.8	635.9	1008.6	1382.5	1756.7	1943.8
$\hat{\beta}_{\gamma}^{\text{MML}}$	1.3	31.8	70.5	127.8	227.7	527.4	714.3	794.3	820.3	834.2
$\hat{\beta}_{\gamma}^{\text{CML}}$	0.3	34.7	80.1	147.6	268.1	659.3	944.6	1137.9	1196.4	1029.7
$\hat{\beta}_{\gamma}^{\text{FB}}$	0.2	34.7	79.5	146.5	267.1	658.0	943.5	1136.7	1195.2	1028.4
$\hat{\beta}_{\gamma}^{\text{FBU}}$	307.2	326.8	353.0	387.9	439.8	594.3	750.6	909.7	1068.7	1150.0

Average losses when $c = 25$

q	0	10	25	50	100	300	500	700	900	1000
C_{MML}	1.5	35.3	75.1	137.4	240.4	570.8	814.0	964.9	999.9	999.8
C_{CML}	0.1	36.1	77.5	141.3	247.6	589.7	844.4	1006.3	1004.5	999.8
C_{FB}	0.1	35.6	77.0	141.1	247.1	589.5	844.2	1006.1	1004.6	999.8
C_{FBU}	633.2	573.8	514.5	467.0	456.8	610.5	812.3	1013.8	1222.0	1327.7
AIC	572.8	578.4	586.3	600.3	628.2	735.9	853.4	963.0	1075.3	1132.9
BIC	76.2	91.5	117.4	162.0	246.0	593.8	943.0	1284.2	1629.7	1808.7
$\hat{\beta}_{\gamma}^{\text{MML}}$	0.6	34.2	73.0	133.7	233.4	551.0	781.7	923.1	957.9	961.1
$\hat{\beta}_{\gamma}^{\text{CML}}$	0.1	35.7	76.5	139.5	243.7	577.3	822.6	976.0	962.6	961.1
$\hat{\beta}_{\gamma}^{\text{FB}}$	0.1	35.2	75.9	139.1	243.2	576.9	822.3	975.7	962.8	961.1
$\hat{\beta}_{\gamma}^{\text{FBU}}$	307.3	373.9	400.3	404.4	420.7	585.3	785.0	984.0	1189.0	1292.5

Table 1: Average losses over 1000 replications of the selection rules $\text{MML} = C_{\text{MML}}$, $\text{CML} = C_{\text{CML}}$, $\text{FB} = C_{\text{FB}}$, $\text{FBU} = C_{\text{FBU}}$, AIC and BIC, and of the posterior mean rules $\hat{\beta}_{\gamma}^{\text{MML}}$, $\hat{\beta}_{\gamma}^{\text{CML}}$, $\hat{\beta}_{\gamma}^{\text{FB}}$ and $\hat{\beta}_{\gamma}^{\text{FBU}}$.

superior. In particular, AIC is much worse when q is small and BIC is much worse when q is large, a reflection of their inability to adapt. Finally, Table 1 and Figure 4 reveal that the estimation-after-selection posterior mean rules, $\hat{\beta}_{\gamma}^{\text{MML}}$, $\hat{\beta}_{\gamma}^{\text{CML}}$, $\hat{\beta}_{\gamma}^{\text{FB}}$ and $\hat{\beta}_{\gamma}^{\text{FBU}}$ all uniformly improve on their least squares counterparts C_{MML} , C_{CML} , C_{FB} and C_{FBU} , especially for $c = 5$ when the components of β are on average smaller and shrinkage is more likely to be effective. Not surprisingly, the relative performance of these four posterior mean rules parallel exactly the relative performance of their counterparts.

4 Discussion

When we began this research, we were hopeful that we would find a Fully Bayes selection procedure that was superior to the Empirical Bayes procedure. We were surprised to find that C_{FB} was not as good as C_{MML} and was essentially

equivalent to C_{CML} . In retrospect, the explanation for this may be understood by noting that MML estimate of c incorporates information across all potential models, whereas the CML estimate is based on the single maximized model. Rather than estimate c , the Fully Bayes procedure margins out c out of each model separately via

$$p(\gamma | Y) \propto \pi(\gamma) \int p(Y | c, \gamma) \pi(c) dc. \quad (46)$$

This integration does not incorporate information outside of the γ th model, and so behaves like the implicit CML posterior $p(\gamma | Y) \propto \pi(\gamma | \hat{w})p(Y | \hat{c}, \gamma)$. Thus the similarity between C_{FB} and C_{CML} was to be expected. Note that the prior $\pi(c)$ is not the issue here.

The dominance of the MML procedure over the FB procedure persisted in our comparisons of the conditional posterior mean estimators $\hat{\beta}_{\gamma}^{FB}$, $\hat{\beta}_{\gamma}^{MML}$ and $\hat{\beta}_{\gamma}^{CML}$. However, in principle, one can go further and consider the Bayes rule under squared error loss, namely the unconditional posterior mean of β under the hyperpriors on c and w . This model averaged estimator would incorporate information across all the potential models and would likely dominate the EB counterparts. Unfortunately, when p is large, such an estimate will be prohibitively expensive to calculate. However, an interesting and computable hybrid alternative proposed by Johnstone and Silverman (2004) integrates out c with respect to a prior, but then uses an empirical Bayes estimate of the median w . Insofar as the posterior median approximates the posterior mean, this estimator is a promising EB alternative.

In this paper, we have focused on the orthogonal case which simplifies the computation of both the EB and FB procedures. However, when X is nonorthogonal, MML calculations will no longer be feasible even in moderately sized problems, and so C_{MML} must be ruled out. Although C_{CML} and C_{FB} can still be pursued, it should be noted that finding the posterior mode will no longer be practical in large problems where it is not feasible to evaluate the posteriors of all the 2^p models. In such cases, heuristics such as stepwise methods or stochastic search might be used to restrict attention to a manageable set of models.

Finally, as mentioned Section 1, we have considered only the case where σ^2 is known in order to focus on the contrast between the EB and FB approaches to hyperparameter uncertainty. However, the more realistic and practical case

of unknown σ^2 can be addressed in all these procedures. A straightforward approach for the EB and FB procedures is to set σ^2 equal to a plug-in estimate such as the traditional full model estimate $(Y'Y - SS_p)/(n - p)$ or, in the nonparametric regression case where $n = p$, to use the robust median estimate $\hat{\sigma} = \text{median}(|\hat{\beta}_i|)/0.6745$, Donoho et al. (1995). But for FB procedures, one can do even better by integrating out σ^2 with respect to a prior such as $\pi(\sigma^2) \propto 1/\sigma^2$ as in Zellner and Siow (1980, 1984) and Liang et. al. (2006).

5 References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B.N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.
- Berger, J.O., (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.
- Berger, J.O. and Pericchi, L. (2001) Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection*, vol. 38 of *IMS Lecture Notes – Monograph Series*, (ed. P. Lahiri), pp. 135–193. Institute of Mathematical Statistics.
- Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101, 157–167.
- Chipman, H., George, E.I. and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*, vol. 38 of *IMS Lecture Notes – Monograph Series*, (ed. P. Lahiri), pp. 65–134. Institute of Mathematical Statistics.
- Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Roy. Statist. Soc. Ser. B*, 62, 681–698.
- Clyde, M. and George, E.I. (2004). Model uncertainty. *Statistical Science*, 19 1 81–94.
- Cui, W. (2002). *Variable Selection: Empirical Bayes vs. Fully Bayes*. Ph. D. Dissertation, Department of MSIS, University of Texas at Austin.

- Donoho, D., Johnstone, I., Kerkycharian, G., and Picard, D., (1995). Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.
- Foster, D.P., and George, E.I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*. **22**, 1947–75.
- Geisser, S. (1984). On Prior Distribution for Binary Trials (with discussion). *The American Statistician*, **38**, 4, 244–251.
- George, E.I. and Foster, D.P. (2000). Calibration and Empirical Bayes variable selection. *Biometrika*, **87**, 4, 731–747.
- George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.
- Jeffreys (1967). *Theory of Probability*, Oxford: University Press.
- Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32, 1594–1649.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J.O. (2006). Mixtures of g-priors for Bayesian Variable Selection, ISDS discussion paper 05-12, Duke University.
- Mallows, C.L. (1973). Some Comments on C_p . *Technometrics*. **15**, 661–676.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. **6**, 461–4.
- Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *Annals of Mathematical Statistics*, 42, 385–388.
- Zellner, A. and Siow, A. (1980). Posterior Odds Ratio for Selected Regression Hypotheses, *Bayesian Statistics 1*, eds J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Valencia: University Press.
- Zellner, A. and Siow, A. (1984). *Basic Issues in Econometrics*, Chicago: University of Chicago Press.

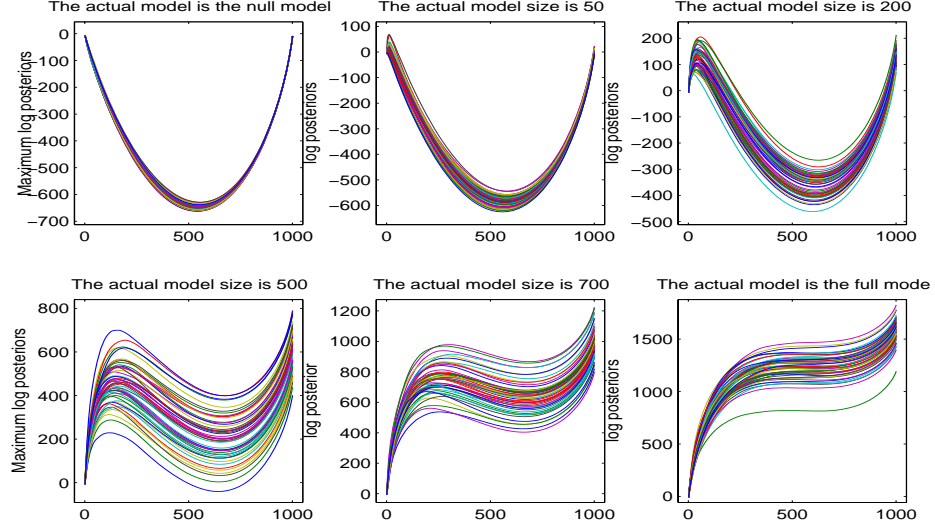


Figure 1: FB maximum log posteriors.

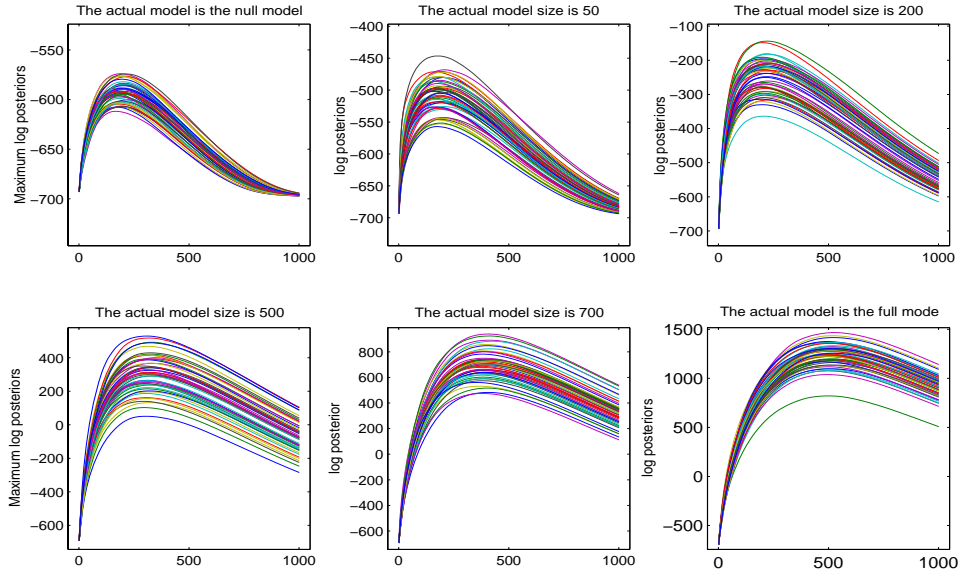


Figure 2: FBU maximum log posteriors.

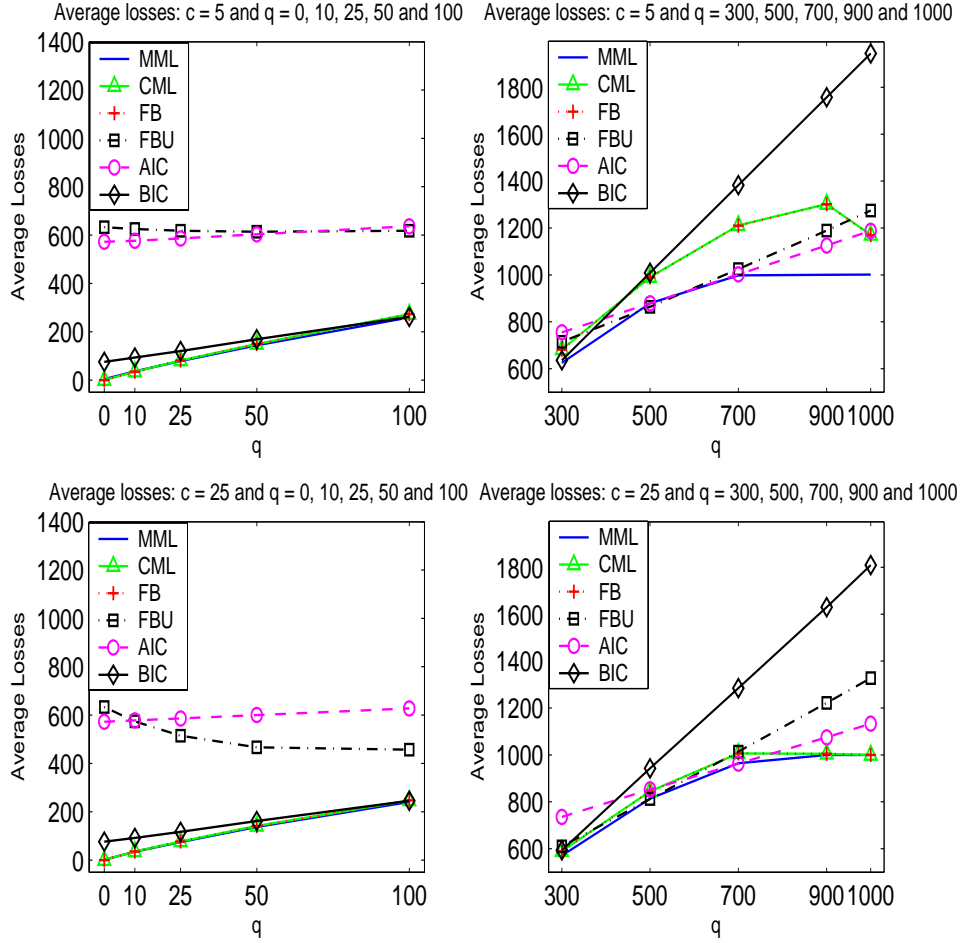


Figure 3: Average losses over 1000 replications of the selection rules $MML = C_{MML}$, $CML = C_{CML}$, $FB = C_{FB}$, $FBU = C_{FBU}$, AIC and BIC .

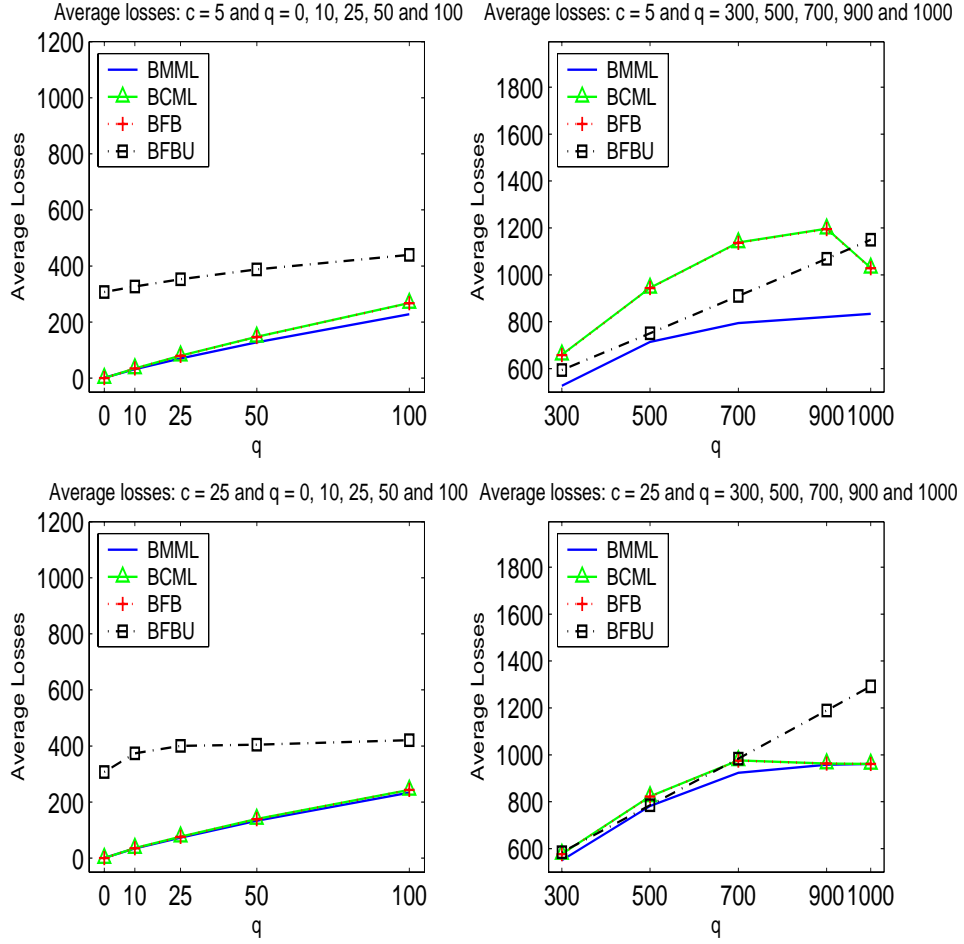


Figure 4: Average losses over 1000 replications of the posterior mean rules $\hat{\beta}_{\gamma}^{MML}$, $\hat{\beta}_{\gamma}^{CML}$, $\hat{\beta}_{\gamma}^{FB}$ and $\hat{\beta}_{\gamma}^{FBU}$.