Efficient Empirical Bayes Variable Selection and Estimation in Linear Models

Ming YUAN and Yi LIN

We propose an empirical Bayes method for variable selection and coefficient estimation in linear regression models. The method is based on a particular hierarchical Bayes formulation, and the empirical Bayes estimator is shown to be closely related to the LASSO estimator. Such a connection allows us to take advantage of the recently developed quick LASSO algorithm to compute the empirical Bayes estimate, and provides a new way to select the tuning parameter in the LASSO method. Unlike previous empirical Bayes variable selection methods, which in most practical situations can be implemented only through a greedy stepwise algorithm, our method gives a global solution efficiently. Simulations and real examples show that the proposed method is very competitive in terms of variable selection, estimation accuracy, and computation speed compared with other variable selection and estimation methods.

KEY WORDS: Hierarchical model; LARS algorithm; LASSO; Model selection; Penalized least squares.

1. INTRODUCTION

We consider the problem of variable selection and coefficient estimation in the common normal linear regression model where we have n observations on a dependent variable Y and p predictors $(x_1, x_2, \ldots, x_p)$, and

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Throughout this article, we center each input variable so that the observed mean is 0 and scale each predictor so that the sample standard deviation is 1.

The underlying notion behind variable selection is that some of the predictors are redundant, and therefore only an unknown subset of the $\boldsymbol{\beta}$ coefficients are nonzero. By effectively identifying the subset of important predictors, variable selection can improve estimation accuracy and enhance model interpretability. Classical variable selection methods, such as C_p, the Akaike introduction criterion, and the Bayes information criterion, choose among possible models using penalized sum of squares criteria, with the penalty being a constant multiple of the model dimension. George and Foster (2000) showed that these criteria correspond to a hierarchical Bayes model selection procedure under a particular class of priors. This gives a new perspective on various earlier model selection methods and puts them in a unified framework. The hierarchical Bayes formulation puts a prior on the model space, then puts a prior on the coefficients given the model. This approach is conceptually attractive. George and Foster (2000) proposed estimating the hyperparameters of the hierarchical Bayes formulation with a marginal maximum likelihood criterion or a conditional maximum likelihood (CML) criterion. The resulting empirical Bayes criterion uses an adaptive dimensionality penalty and compares favorably with the penalized least squares criteria with fixed dimensionality penalty. However, even after the hyperparameters are estimated, the resulting model selection criterion must be evaluated on each candidate model to select the best model. This is impractical for even a moderate number of predictors, because the number of candidate models grows exponentially as the number of predictors increases. In practice, this type of method is implemented in a stepwise fashion through forward selection or backward elimination. In doing so, one contents oneself with the locally optimal solution instead of the globally optimal solution.

A number of other variable selection methods have been introduced in recent years (George and McCulloch 1993; Foster and George 1994; Breiman 1995; Tibshirani 1996; Fan and Li 2001; Shen and Ye 2002; Efron, Johnston, Hastie, and Tibshirani 2004). In particular, Efron et al. (2004) proposed an effective variable selection algorithm, LARS (least-angle regression), that is extremely fast, and showed that with slight modification the LARS algorithm can be used to efficiently compute the popular LASSO estimate for variable selection, defined as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \Big(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum |\beta_i| \Big), \tag{2}$$

where $\lambda > 0$ is a regularization parameter. By using the L_1 penalty, minimizing (2) yields a sparse estimate of $\boldsymbol{\beta}$ if λ is chosen appropriately. Consequently, a submodel of (1) containing only the covariates corresponding to the nonzero components in $\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\lambda)$ is selected as the final model. The LARS algorithm computes the whole path of the LASSO with a computational load in the same magnitude as the ordinary least squares. Therefore, computation is extremely fast, which facilitates the choice of the tuning parameter with criteria such as C_p or generalized cross-validation (GCV).

In this article we adopt a hierarchical Bayes framework similar to that of George and McCulloch (1993) and George and Foster (2000), but with new prior specifications. We show that the resulting empirical Bayes estimator is closely related to the LASSO estimator and is quickly computable. We introduce a method for choosing the hyperparameters, which in turn leads to an alternative method for choosing the tuning parameter in the LASSO. Unlike earlier methods, including that of George and Foster (2000) and the LASSO tuned with C_p, where the error variance σ^2 is assumed known or fixed at the estimate from the saturated model, in our method it is estimated together with the other parameters. Therefore, our method potentially can be

Ming Yuan is Assistant Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (E-mail: *myuan@isye.gatech.edu*). Yi Lin is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: *yilin@stat.wisc.edu*). Yuan's research was supported in part by National Science Foundation grant DMS-00-72292; Lin's research was supported in part by National Science Foundation grant DMS-01-34987. The authors thank the editor, the associate editor, and two anonymous referees for comments that greatly improved the manuscript.

Journal of the American Statistical Association
December 2005, Vol. 100, No. 472, Theory and Methods
DOI 10.1198/016214505000000367

used in situations when the dimension is larger than the sample size.

The rest of the article is structured as follows. In Section 2 we introduce a hierarchical Bayes formulation for variable selection. In Section 3 we present an analytic approximation to the posterior probabilities in the Bayesian formulation that turns out to be connected to the LASSO estimate. We further justify this connection theoretically under the condition of orthogonal design in Section 4. In Section 5 we propose a method for choosing the hyperparameters. In Section 6 we conduct simulation studies to compare our method with some related model selection methods. We illustrate the performance of our method on several real datasets in Section 7, and provide a summary in Section 8.

2. HIERARCHICAL MODEL FORMULATION

A hierarchical model formulation for variable selection in linear models consists of the following three main ingredients:

1. A prior probability, $P(\mathcal{M})$, for each candidate model $\mathcal{M}$
2. A prior, $P(\theta_{\mathcal{M}}|\mathcal{M})$, for parameter $\theta_{\mathcal{M}}$ associated with model $\mathcal{M}$
3. A data-generating mechanism conditional on $(\mathcal{M}, \theta_{\mathcal{M}})$, $P(Y|\mathcal{M}, \theta_{\mathcal{M}})$.

Once these three components are specified, one can combine data and priors to form the posterior,

$$P(\mathcal{M}|\mathbf{Y}) = \frac{P(\mathcal{M}) \int P(\mathbf{Y}|\mathcal{M}, \theta_{\mathcal{M}}) P(\theta_{\mathcal{M}}|\mathcal{M})\, d\theta_{\mathcal{M}}}{\sum_{\mathcal{M}'} \int P(\mathbf{Y}|\mathcal{M}', \theta_{\mathcal{M}'}) P(\theta_{\mathcal{M}'}|\mathcal{M}')\, d\theta_{\mathcal{M}'} P(\mathcal{M}')}. \tag{3}$$

We begin by indexing each candidate model with one binary vector, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$. An element γ_i takes value 0 or 1 depending on whether or not the ith predictor is excluded from the model. Adopting this notation, under model $\boldsymbol{\gamma}$, (1) can be written as

$$\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\beta} \sim \mathrm{N}\big(\mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mathbf{I}_{(n)}\big), \tag{4}$$

where subscript "$\boldsymbol{\gamma}$" indicates that only those columns or elements with the corresponding $\boldsymbol{\gamma}$ element of 1 are included. Notice that $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is of dimension $|\boldsymbol{\gamma}|$, where $|\boldsymbol{\gamma}|$ denotes $\sum \gamma_i$.

Now we specify the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. By the definition of $\boldsymbol{\gamma}$, it is natural to force $\beta_i = 0$ if $\gamma_i = 0$. But if $\gamma_i = 1$, then we give a double-exponential prior for β_i, that is,

$$\beta_i|\gamma_i = (1 - \gamma_i)\delta(0) + \gamma_i DE(0, \tau), \qquad j = 1, \ldots, p, \tag{5}$$

where $DE(0, \tau)$ has density function $\tau \exp(-\tau|x|)/2$. In contrast with the commonly used normal prior $\beta_i|\gamma_i = 1 \sim \mathrm{N}(0, \tau^2)$, the double-exponential prior can better accommodate large regression coefficients because of its heavier tail probability. The double-exponential prior can also be presented as a two-level hierarchical model (Andrews and Mallows 1974). At the first level, the regression coefficient β_i is assumed to follow $\beta_i|\gamma_i = 1 \sim \mathrm{N}(0, \eta_i)$. At the second level, an exponential prior is assumed for η's, $\eta_i \sim \exp(\tau^2/2)$. From this, we can see that the double-exponential prior introduces different variance parameters for different regression coefficients. In a wavelet setup, Johnstone and Silverman (2005) argued that the double-exponential prior can achieve the adaptive minimax convergence rates not obtainable using normal priors.

We remark that there is another approach to variable selection in linear models. Instead of putting a degenerate prior on β_j for j's with $\gamma_j = 0$, one can put a prior on the full set of β's and then exclude variables with small effects. This smoother alternative to our formulation and has been adopted by George and McCulloch (1993), among others.

For $\boldsymbol{\gamma}$, a widely used prior is $P(\boldsymbol{\gamma}) = q^{|\boldsymbol{\gamma}|}(1 - q)^{p - |\boldsymbol{\gamma}|}$, with a prespecified q. This prior assumes that each predictor enters the model independently with a prior probability q, whether or not the predictors are correlated. The prior models the prior information on the model sizes but does not distinguish models with the same size. However, it is often the case that highly correlated predictors are to be avoided simply because those predictors are providing similar information on the response. To achieve this, we propose the following prior for $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma}) \propto q^{|\boldsymbol{\gamma}|}(1 - q)^{p - |\boldsymbol{\gamma}|} \sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}})}, \tag{6}$$

where $\det(\mathbf{X}'_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}}) = 1$ if $|\boldsymbol{\gamma}| = 0$. Note that when the correlation between two covariates goes to 1, the prior described in (6) converges to a prior that allows only one of the two variables in the model. To better appreciate the effect of correlation between predictors in our prior specification, consider the conditional prior odds ratio for $\gamma_j = 1$,

$$\frac{P(\gamma_j = 1|\boldsymbol{\gamma}^{[-j]})}{P(\gamma_j = 0|\boldsymbol{\gamma}^{[-j]})} = \frac{q}{1 - q} \sqrt{\frac{\det(\mathbf{X}'_{\boldsymbol{\gamma}^{[-j]}, \gamma_j = 1} \mathbf{X}_{\boldsymbol{\gamma}^{[-j]}, \gamma_j = 1})}{\det(\mathbf{X}'_{\boldsymbol{\gamma}^{[-j]}, \gamma_j = 0} \mathbf{X}_{\boldsymbol{\gamma}^{[-j]}, \gamma_j = 0})}}, \tag{7}$$

where superscript "$[-j]$" indicates that the jth component is removed. If X_j is highly correlated with the current covariates, $X_{\boldsymbol{\gamma}^{[-j]}, \gamma_j = 0}$, then the second factor of the right side of (7) will be small. Therefore, it is more likely that X_j will be removed from the full model. This is desirable, because X_j does not contain much "additional" information.

Our Bayesian formulation consists of (4), (5), and (6). Three parameters need to be specified for this formulation, namely q, τ, and σ^2. From a hierarchical Bayesian standpoint, one can either use prespecified values or put a higher level prior for them. Both of these approaches require human expertise. To avoid the need for expert information, we take the empirical Bayes approach and use an automatic default prior parameter choice. The automatic choice of the hyperparameters is introduced in Section 5.

With our formulation, the joint distribution $P(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{Y})$ is

$$\begin{aligned} &P(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{Y}) \\ &\quad \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}} \\ &\qquad \times \exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{i \in \boldsymbol{\gamma}} |\beta_i|}{2\sigma^2}\right) (1 - q)^p w^{|\boldsymbol{\gamma}|}. \end{aligned}$$

Therefore,

$$\begin{aligned} P(\boldsymbol{\gamma}|\mathbf{Y}) &= C(\mathbf{Y}) w^{|\boldsymbol{\gamma}|} \\ &\quad \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}} \\ &\quad \times \exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{i \in \boldsymbol{\gamma}} |\beta_i|}{2\sigma^2}\right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \end{aligned} \tag{8}$$

where

$$w = \left(\frac{q}{1-q}\frac{\tau}{2}\sqrt{2\pi\sigma^2}\right),$$

$\lambda = 2\sigma^2\tau$, and $C(\mathbf{Y})$ is a constant not depending on $\boldsymbol{\gamma}$. We pick the model $\boldsymbol{\gamma}$ that maximizes $P(\boldsymbol{\gamma}|\mathbf{Y})$. In principle, exact evaluation of the posterior probability $P(\boldsymbol{\gamma}|\mathbf{Y})$ could be obtained; however, this task cannot be performed in closed form. To compute the high-dimensional integrals involved in $P(\boldsymbol{\gamma}|\mathbf{Y})$, analytical or numerical approximation methods are needed.

3. POSTERIOR ANALYSIS

The major difficulty in posterior inference for our Bayesian model comes from the high-dimensional integration in (8). Because no analytically tractable solution to this integral exists in general, we have to use approximations. Because the posterior probability is expected to spread over a large number of possible models, it is not possible to construct analytical approximations that do uniformly well for all candidate models. We propose to focus on a subset of candidate models containing the model with the highest posterior probability, whose posterior probabilities can be approximated very well.

Let

$$\boldsymbol{\beta}^*_{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left(\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i|\right).$$

Denote $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \boldsymbol{\beta}^*_{\boldsymbol{\gamma}} + \mathbf{u}$. We can rewrite (8) as

$$\begin{aligned}
P(\boldsymbol{\gamma}|\mathbf{Y}) &= C(\mathbf{Y})w^{|\boldsymbol{\gamma}|}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}}\\
&\quad\times\exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i|}{2\sigma^2}\right)d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\\
&= C(\mathbf{Y})w^{|\boldsymbol{\gamma}|}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}}\\
&\quad\times\exp\left(-\left(\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\right.\right.\\
&\qquad\left.\left.+\lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta^*_i + u_i| - |\beta^*_i|)\right)\Big/(2\sigma^2)\right)d\mathbf{u}\\
&\quad\times\exp\left(-\frac{\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}(\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i|)}{2\sigma^2}\right),
\end{aligned} \tag{9}$$

where $\tilde{\mathbf{Y}}_{\boldsymbol{\gamma}} = \mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$, and hereafter we omit the subscript "$\boldsymbol{\gamma}$" if no confusion occurs. Our main task is to evaluate

$$\begin{aligned}
&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}}\\
&\quad\times\exp\left(-\left(\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\right.\right.\\
&\qquad\left.\left.+\lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta^*_i + u_i| - |\beta^*_i|)\right)\Big/(2\sigma^2)\right)d\mathbf{u}.
\end{aligned} \tag{10}$$

Define

$$f(\mathbf{u}) \equiv \frac{\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u} + \lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta^*_i + u_i| - |\beta^*_i|)}{\sigma^2}. \tag{11}$$

Note that the definition of f depends on $\boldsymbol{\gamma}$ implicitly, because it has a $|\boldsymbol{\gamma}|$-dimensional argument. Hereafter we omit this dependence for notational convenience. From the definition of $\mathbf{u}$, we see that $f(\mathbf{u})$ is minimized at $\mathbf{u}^* = \mathbf{0}$.

Now we consider the following two types of models separately.

Definition 1. For a dataset $(\mathbf{X}, \mathbf{Y})$ and a given regularization parameter λ,

(a) a model $\boldsymbol{\gamma}$ is called *regular* if and only if $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ does not contain 0's or $|\boldsymbol{\gamma}| = 0$ and

(b) a model $\boldsymbol{\gamma}$ is called *nonregular* if $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ contains at least one zero component.

3.1 Regular Models

For regular models, $f(\mathbf{u})$ is differentiable at $\mathbf{u} = \mathbf{u}^*$, and

$$\left.\frac{\partial^2 f(\mathbf{u})}{\partial\mathbf{u}\,\partial\mathbf{u}^T}\right|_{\mathbf{u}=\mathbf{u}^*} = \frac{1}{\sigma^2}\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}. \tag{12}$$

Applying the Laplace approximation to (10), we get, for sample size n large enough,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}}\\
&\quad\times\exp\left(-\left(\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\right.\right.\\
&\qquad\left.\left.+\lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta^*_i + u_i| - |\beta^*_i|)\right)\Big/(2\sigma^2)\right)d\mathbf{u} \approx 1.
\end{aligned} \tag{13}$$

It is worth pointing out that we implicitly assume the nonsingularity of $\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}$ when using the Laplace approximation. This should not be a loss of generality, however. The models that do not satisfy this condition are not of interest from a variable selection standpoint and have been assigned prior probability 0 [see (6)]. Therefore, the posterior probability for these models is always 0.

Combining (9) and (13), for sample size n large enough, we have

$$\begin{aligned}
&P(\boldsymbol{\gamma}|\mathbf{Y})\\
&\quad\approx C(\mathbf{Y})w^{|\boldsymbol{\gamma}|}\\
&\qquad\times\exp\left(-\frac{\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}(\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i|)}{2\sigma^2}\right).
\end{aligned} \tag{14}$$

Although (14) is derived for large sample sizes, we find this approximation to be fairly accurate even for small sample sizes. To elaborate on this, we simulated a dataset with $p = 8$ covariates and $n = 20$ observations. The regression coefficients are generated independently from $\mathrm{N}(0, 2^2)$, and the regression

Table 1. Maximum Approximation Error of (14) for Different Sample Sizes

n	Minimum	First quartile	Median	Mean	Third quartile	Maximum
20	6.05%	9.37%	11.02%	11.42%	12.66%	23.89%
50	3.85%	5.82%	6.76%	7.00%	7.72%	14.25%
100	2.68%	4.04%	4.70%	4.83%	5.44%	9.34%

noise follows $N(0, 3^2)$. The design matrix is generated by orthogonalizing independent standard normal variables. The left side of (13) represents the ratio of the exact posterior probability to the corresponding approximation given by (14) for each regular model given $\lambda > 0$. It can be factorized into univariate integrals and thus is readily computable under orthogonal design. For every $\lambda > 0$, we computed this ratio for every regular model and recorded the ratio that differs most from 1 over all regular models. For a typical dataset, the largest discrepancy between the recorded ratio and 1 taken over all $\lambda > 0$ is only about 10%. We then repeated the experiment for three different sample sizes, $n = 20, 50$, and 100. Table 1 presents the summary statistics of the largest discrepancies taken over 100 runs. We can see that (14) is quite accurate for sample sizes as small as 20.

3.2 Nonregular Models

Although (14) provides a computationally efficient approximation to $P(\boldsymbol{\gamma}|\mathbf{Y})$ for regular models, it does not apply to nonregular models, because for these models, $f(\mathbf{u})$ is not differentiable at $\mathbf{u} = \mathbf{u}^*$. However, in what follows we show that in our model selection procedure, we can concentrate on the regular models.

It is beneficial to exclude complex models that do not receive more support from the data than their simpler counterparts. For this reason, we compare a nonregular model $\boldsymbol{\gamma}$ with a regular submodel of $\boldsymbol{\gamma}$. Without loss of generality, assume that $\boldsymbol{\gamma}$ is of the form $(1, \ldots, 1, 0, \ldots, 0)$, where the first $|\boldsymbol{\gamma}|$ components are 1's, and that only the first s components of the $|\boldsymbol{\gamma}|$-dimensional vector $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ are nonzero. By the definition of nonregular models, we have $s < |\boldsymbol{\gamma}|$. Let $\boldsymbol{\gamma}^*$ be the p-dimensional binary vector representing a submodel of $\boldsymbol{\gamma}$ with only the first s elements being 1. Our task here is to compare $P(\boldsymbol{\gamma}|\mathbf{Y})$ and $P(\boldsymbol{\gamma}^*|\mathbf{Y})$. Because $f(\mathbf{u})$ is minimized at $\mathbf{u} = \mathbf{0}$, for any $i \le s$, we have

$$\left.\frac{\partial f}{\partial u_i}\right|_{\mathbf{u}=\mathbf{0}} = 0,$$

which leads to

$$2\tilde{\mathbf{Y}}' X_i = \lambda \operatorname{sign}(\beta^*_{\boldsymbol{\gamma},i}) \quad \text{if } i \le s. \tag{15}$$

In contrast, for $s < i \le |\boldsymbol{\gamma}|$, the ith component of $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ is 0, and we have

$$\left.\frac{\partial f}{\partial u_i}\right|_{u_i=0^+;u_j=0,\forall j\neq i} \ge 0 \qquad \text{and} \qquad \left.\frac{\partial f}{\partial u_i}\right|_{u_i=0^-;u_j=0,\forall j\neq i} \le 0.$$

This implies that

$$|2\tilde{\mathbf{Y}}'\mathbf{X}_i| \le \lambda \qquad \text{and} \qquad \beta^*_{\boldsymbol{\gamma},i} = 0 \quad \text{if } s < i \le |\boldsymbol{\gamma}|. \tag{16}$$

By (15) and (16), from simple calculations, we get

$$\begin{aligned}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} &\frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}} \\ &\times \exp\Bigg(-\Bigg(\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u} \\ &\qquad + \lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta^*_i + u_i| - |\beta^*_i|)\Bigg)\Big/(2\sigma^2)\Bigg)\,d\mathbf{u} \\ < \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} &\frac{\sqrt{\det(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})}}{(\sqrt{2\pi\sigma^2})^{|\boldsymbol{\gamma}|}}\exp\left(-\frac{\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2}{2\sigma^2}\right)d\mathbf{u} = 1.\end{aligned}$$

Thus

$$\begin{aligned}&P(\boldsymbol{\gamma}|\mathbf{Y}) \\ &\quad < C(\mathbf{Y})w^{|\boldsymbol{\gamma}|} \\ &\qquad \times \exp\left(-\frac{\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}(\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i|)}{2\sigma^2}\right).\end{aligned} \tag{17}$$

Now, because $\tilde{\mathbf{Y}}_{\boldsymbol{\gamma}} = \tilde{\mathbf{Y}}_{\boldsymbol{\gamma}^*}$ and $\beta^*_{\boldsymbol{\gamma},i} = \beta^*_{\boldsymbol{\gamma}^*,i}$ for any $i \le s$, applying (14) to the regular model $\boldsymbol{\gamma}^*$ and (17) to the nonregular model $\boldsymbol{\gamma}$, we conclude that, asymptotically,

$$\frac{P(\boldsymbol{\gamma}|\mathbf{Y})}{P(\boldsymbol{\gamma}^*|\mathbf{Y})} \le w^{|\boldsymbol{\gamma}|-s}.$$

If $w \le 1$, the data do not give more support to the bigger model $\boldsymbol{\gamma}$ than to $\boldsymbol{\gamma}^*$, and thus we would pick $\boldsymbol{\gamma}^*$. Consequently, we can avoid computing $P(\boldsymbol{\gamma}|\mathbf{Y})$ for nonregular model $\boldsymbol{\gamma}$.

4. CONNECTION BETWEEN THE LASSO ALGORITHM AND THE BAYESIAN FRAMEWORK

Summarizing the foregoing analysis, we find that if w is set to 1, then the following assumptions hold:

1. To search for the model with the highest posterior probability, we can concentrate on the regular models.
2. For regular models, the posterior probability, $P(\boldsymbol{\gamma}|\mathbf{Y})$, can be approximated by $C(\mathbf{Y})\exp[-h(\boldsymbol{\gamma})/(2\sigma^2)]$, where

$$h(\boldsymbol{\gamma}) = \min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left(\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i|\right).$$

In general, these conclusions are good approximations. For the special case of orthogonal design matrix $\mathbf{X}$, they can be proved rigorously.

Theorem 1. Suppose that $w = 1$. Under orthogonal design, that is, $\mathbf{X}'\mathbf{X} = (n-1)\mathbf{I}_p$, the following results hold:

a. If model $\boldsymbol{\gamma}$ is nonregular, then there exists a $\boldsymbol{\gamma}^*$ such that $P(\boldsymbol{\gamma}|\mathbf{Y}) < P(\boldsymbol{\gamma}^*|\mathbf{Y})$.

b. Suppose that $\lambda = o(\sqrt{n})$. If model $\boldsymbol{\gamma}$ is regular, then, as $n \to \infty$,

$$P(\boldsymbol{\gamma}|\mathbf{Y}) \sim C(\mathbf{Y}) \exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{i\in\boldsymbol{\gamma}} |\beta^*_i|}{2\sigma^2}\right).$$

We note the $(n-1)$ factor in the condition $\mathbf{X}'\mathbf{X} = (n-1)\mathbf{I}_p$ is just to conform to our convention we stated at the beginning of the article that the data be scaled to have sample standard deviation 1.

By parts a and b, we can now focus on searching for the regular model $\boldsymbol{\gamma}$ with the smallest $h(\boldsymbol{\gamma})$. A straightforward search involves going through each of the large number of candidate models to identify regular models and to minimize h over all of the regular models, which would be computationally very demanding. Fortunately, such a search is not necessary, and the following proposition provides the key to a simple and explicit recipe for finding a regular model that minimizes $h(\boldsymbol{\gamma})$. The proof of the proposition is relegated to the Appendix.

Proposition 1. Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^p |\beta_i|)$, and let model $\widehat{\boldsymbol{\gamma}}$ be such that $\widehat{\gamma}_i = I(\widehat{\beta}_i \neq 0)$, where $I(\cdot)$ is the indicator function. Then $\widehat{\boldsymbol{\gamma}}$ is the regular model that minimizes $h(\boldsymbol{\gamma})$.

Interestingly, $\widehat{\boldsymbol{\gamma}}$ is exactly the same model selected by the LASSO algorithm. In other words, the model selected by the LASSO algorithm has the highest posterior probability under our Bayesian model (6) with $w = 1$. Therefore, we can use the LASSO algorithm to select a model with approximately the largest posterior probability when $w = 1$. The LASSO algorithm also gives the maximum a posteriori estimate of the regression coefficients for the selected model at the same time. Thus, if our goal is to select a model and estimate the regression coefficient, then we can use the LASSO algorithm to fulfill the task. This equivalence allows us to take advantage of the recently developed fast LASSO algorithm to compute the solution for our Bayesian formulation. The connection with the LASSO estimator also highlights a distinction between our empirical Bayes methods and earlier proposals, such as that of George and Foster (2000) which can be implemented only in a stepwise fashion in most practical situations. Using the LASSO algorithm to calculate the Bayesian solution would save tremendous computational effort and make our procedure suitable for large datasets with high dimensionality.

The close relationship also gives a new Bayesian interpretation to the LASSO algorithm. Tibshirani (1996) mentioned that the LASSO algorithm has another Bayesian interpretation with an independent double-exponential prior on each regression coefficient. Tibshirani's formulation is somehow less natural as a Bayesian variable selection procedure, because it puts prior probability 1 on the full model. Consequently, the corresponding posterior probability for the full model will also be 1 even if the posterior modal estimates of some regression coefficients are 0.

Although setting $w = 1$ substantially eases the computational burden, it potentially may incur loss of efficiency in terms of prediction accuracy. However, we found that this potential loss of efficiency is usually small. To illustrate, we conducted a small experiment where $y_i = \beta_i + \epsilon_i$, $i = 1, \ldots, 500$. The noise ϵ_i follows a standard normal distribution. Notice that any linear regression problem with orthogonal design can be transformed into this form. In our experiment, the first 50 β_i's are generated from N(0, 1), and the rest are set at 0. This orthogonal design allows us to restrict our attention on a sequence of 500 submodels instead of all 2^{500} submodels and perform exact posterior analysis. For each of 100 equally spaced $\lambda \in [0, \max|y_i|]$ and any $w \geq 0$, we computed the model with the highest posterior probability together with its associated coefficient estimate $\widehat{\boldsymbol{\beta}}_{w,\lambda}$. Consequently, we can compute $\widehat{\boldsymbol{\beta}}_{\text{opt}} = \arg\min_{w,\lambda} \|\widehat{\boldsymbol{\beta}}_{w,\lambda} - \boldsymbol{\beta}\|^2$ and $\widehat{\boldsymbol{\beta}}_{\text{opt},w=1} = \arg\min_{\lambda} \|\widehat{\boldsymbol{\beta}}_{1,\lambda} - \boldsymbol{\beta}\|^2$. The summary statistics of the relative estimation efficiency $\|\widehat{\boldsymbol{\beta}}_{\text{opt}} - \boldsymbol{\beta}\|^2 / \|\widehat{\boldsymbol{\beta}}_{\text{opt},w=1} - \boldsymbol{\beta}\|^2$ over 100 runs are reported in Table 2.

Table 2. Relative Efficiency by Forcing $w = 1$

Minimum	First quartile	Median	Mean	Third quartile	Maximum
66.67%	92.66%	97.32%	95.01%	99.36%	100.00%

As Table 2 clearly indicates, the loss of efficiency caused by forcing $w = 1$ is small. Therefore, it is reasonable to set w to 1, given the great computational advantage that setting $w = 1$ brings about.

5. PRIOR ELICITATION

After setting $w = 1$, we still need to specify σ^2 and λ. The later is exactly the tuning parameter selection problem faced by the LASSO algorithm. Tibshirani (1996) proposed a GCV score for selecting λ. In what follows, we adopt an empirical Bayesian approach for selecting both σ^2 and λ.

From an empirical Bayesian standpoint, one could choose σ^2 and λ by maximizing the marginal likelihood

$$f(\mathbf{Y}|\sigma^2, \lambda) = \sum_{\boldsymbol{\gamma}} \int_{-\infty}^{\infty} P(\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}.$$

This can be implemented when the number of variables is small. But in situations where the number of variables is moderately large, the summation is over a large number of items and is not practical for large datasets. In such situations we follow an approach related to the CML approach proposed by George and Foster (2000). The conditional density of $\mathbf{Y}$ given a model $\boldsymbol{\gamma}$ is

$$\begin{aligned} &f(\mathbf{Y}|\boldsymbol{\gamma}, \sigma^2, \lambda) \\ &\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2}{2\sigma^2}\right) \\ &\qquad \times \left(\frac{\lambda}{4\sigma^2}\right)^{|\boldsymbol{\gamma}|} \exp\left(-\frac{\lambda \sum_{i\in\boldsymbol{\gamma}} |\beta_i|}{2\sigma^2}\right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \end{aligned}$$

For a given λ, denote the selected model by $\widehat{\boldsymbol{\gamma}}_\lambda$. We choose σ^2 and λ as the maximizer of $f(\mathbf{Y}|\widehat{\boldsymbol{\gamma}}_\lambda, \sigma^2, \lambda)$. Strictly speaking, $f(\mathbf{Y}|\widehat{\boldsymbol{\gamma}}_\lambda, \sigma^2, \lambda)$ is neither a likelihood nor a conditional likelihood, and we denote the resulting criterion by CML only because of its similarity to the approach of George and Foster (2000). Because $\widehat{\boldsymbol{\gamma}}_\lambda$ is regular, we can use (9) and (13) to approximate $f(\mathbf{Y}|\widehat{\boldsymbol{\gamma}}_\lambda, \sigma^2, \lambda)$,

$$\begin{aligned} &f(\mathbf{Y}|\widehat{\boldsymbol{\gamma}}_\lambda, \sigma^2, \lambda) \\ &\quad \approx \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-|\widehat{\boldsymbol{\gamma}}_\lambda|} \left(\frac{\lambda}{4\sigma^2}\right)^{|\widehat{\boldsymbol{\gamma}}_\lambda|} \left(\det\left(\mathbf{X}'_{\widehat{\boldsymbol{\gamma}}_\lambda}\mathbf{X}_{\widehat{\boldsymbol{\gamma}}_\lambda}\right)\right)^{-1/2} \end{aligned}$$

$$\times \exp\left(-\frac{\min_{\boldsymbol{\beta}}(\|\mathbf{Y}-\mathbf{X}_{\widehat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\beta}_{\widehat{\boldsymbol{\gamma}}_\lambda}\|^2+\lambda\sum_{i\in\widehat{\boldsymbol{\gamma}}_\lambda}|\beta_i|)}{2\sigma^2}\right)$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-|\widehat{\boldsymbol{\gamma}}_\lambda|}\left(\frac{\lambda}{4\sigma^2}\right)^{|\widehat{\boldsymbol{\gamma}}_\lambda|}\left(\det\left(\mathbf{X}'_{\widehat{\boldsymbol{\gamma}}_\lambda}\mathbf{X}_{\widehat{\boldsymbol{\gamma}}_\lambda}\right)\right)^{-1/2}$$

$$\times \exp\left(-\frac{\min_{\boldsymbol{\beta}}(\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\sum_{i=1}^{p}|\beta_i|)}{2\sigma^2}\right).$$

Simple calculations show that this is equivalent to choosing λ by minimizing

$$\mathrm{CML}(\lambda)$$

$$\equiv (n+|\widehat{\boldsymbol{\gamma}}_\lambda|)\left[\ln\left(\frac{\min_{\boldsymbol{\beta}}(\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\sum_{i=1}^{p}|\beta_i|)}{n+|\widehat{\boldsymbol{\gamma}}_\lambda|}\right)+1\right]$$

$$+\ln\left(\det\left(\mathbf{X}'_{\widehat{\boldsymbol{\gamma}}_\lambda}\mathbf{X}_{\widehat{\boldsymbol{\gamma}}_\lambda}\right)\right)-2|\widehat{\boldsymbol{\gamma}}_\lambda|\ln(\sqrt{2\pi}\lambda/4),$$

and the estimate of σ^2 is $\min_{\boldsymbol{\beta}}(\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\sum_{i=1}^{p}|\beta_i|)/(n+|\widehat{\boldsymbol{\gamma}}_\lambda|)$.

Remark 1. Both C_p used by Efron et al. (2004) and the empirical Bayes approach previously proposed by George and Foster (2000) need to estimate σ^2 by fitting the full model and thus can be applied only in the situation where $p<n$. Our proposed approach avoids this problem.

Remark 2. It is interesting to notice that the derivation for CML does not work for the Bayesian interpretation given by Tibshirani (1996) mentioned at the end of Section 4, because the approximation applies only to regular models. However, it is tempting to maximize $f(\mathbf{Y}|\widehat{\boldsymbol{\beta}}_\lambda,\widehat{\boldsymbol{\gamma}}_\lambda,\sigma^2,\lambda)$ instead. Unfortunately, this leads to a criterion,

$$\min_{\boldsymbol{\beta}}\left(\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\sum_{i=1}^{p}|\beta_i|\right),$$

that is trivially minimized at $\lambda=0$.

6. SIMULATIONS

In this section we compare the proposed empirical Bayes procedure with several other popular approaches for variable selection and estimation. The methods compared include:

- EBC, our approximate empirical Bayes estimate with hyperparameters selected by CML
- LCP, the LASSO with λ selected by C_p
- LGCV, the LASSO with λ selected by GCV
- GFF, the empirical Bayes approach proposed George and Foster (2000) and implemented in a forward-selection fashion. George and Foster proposed a conditional maximum likelihood method for choosing the hyperparameters in their Bayes formulation.

We compare these methods in terms of the size of selected models, model error, and the computation time on a Pentium III 750-M computer. All simulations were conducted using R. The path of the LASSO estimate was computed using the R package LARS. The model error of an estimate $\widehat{\boldsymbol{\beta}}$ is given by

$$ME(\widehat{\boldsymbol{\beta}})=(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{V}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}),$$

where $\mathbf{V}=E(\mathbf{X}'\mathbf{X})$ is the population covariance matrix of $\mathbf{X}$. The models in our first simulation example were also used by Tibshirani (1996).

Example 1. Consider the following four models:

I. $\boldsymbol{\beta}=(3,1.5,0,0,2,0,0,0)'$ and $\sigma=3$. The correlation between $\mathbf{X}_{\cdot i}$ and $\mathbf{X}_{\cdot j}$ is $\rho^{|i-j|}$ with $\rho=.5$.

II. Same as model I except that $\beta_j=.85$ for all j.

III. Same setup as before, but with $\boldsymbol{\beta}=(5,0,0,0,0,0,0,0)'$ and $\sigma=2$.

IV. A total of 40 correlated predictors are considered. $x_{ij}=z_{ij}+w_i$, where z_{ij} and w_i are independent standard normal random variables. The true regression coefficients are 2 for the first 20 predictors and 0 for the other predictors.

For models I–III, 200 datasets with sample size 20 were generated. For the model IV, 200 datasets with sample size 100 were generated.

Table 3 gives the means and standard errors over the 200 simulated datasets. The table indicates that EBC tends to select models with relatively smaller size than the other methods. To check see whether the choice of sparse models comes at a sacrifice of prediction accuracy, we provide a pairwise prediction accuracy comparison between EBC and the other methods

Table 3. Comparison of the Simulated Datasets

	EBC		*LCP*		*LGCV*		*GFF*	
Model I								
ME	3.99	(.24)	5.19	(.37)	4.52	(.26)	6.37	(.34)
Size	5.14	(.08)	5.29	(.11)	7.37	(.05)	5.66	(.21)
Time (sec)	.11	(0)	.05	(0)	.23	(0)	.27	(0)
Model II								
ME	4.95	(.23)	5.60	(.32)	4.76	(.25)	6.55	(.33)
Size	5.68	(.08)	5.70	(.10)	7.22	(.06)	6.80	(.18)
Time (sec)	.11	(0)	.05	(0)	.23	(0)	.27	(0)
Model III								
ME	1.19	(.09)	1.81	(.17)	1.70	(.10)	.64	(.13)
Size	4.23	(.09)	4.06	(.16)	7.26	(.05)	1.32	(.09)
Time (sec)	.11	(0)	.05	(0)	.23	(0)	.29	(0)
Model IV								
ME	61.18	(1.05)	80.22	(2.26)	87.18	(1.98)	183.47	(2.66)
Size	25.45	(.16)	27.26	(.32)	33.43	(.22)	6.89	(.09)
Time (sec)	.87	(.01)	.30	(0)	5.01	(.03)	8.96	(.01)

NOTE: Standard errors are in parentheses.

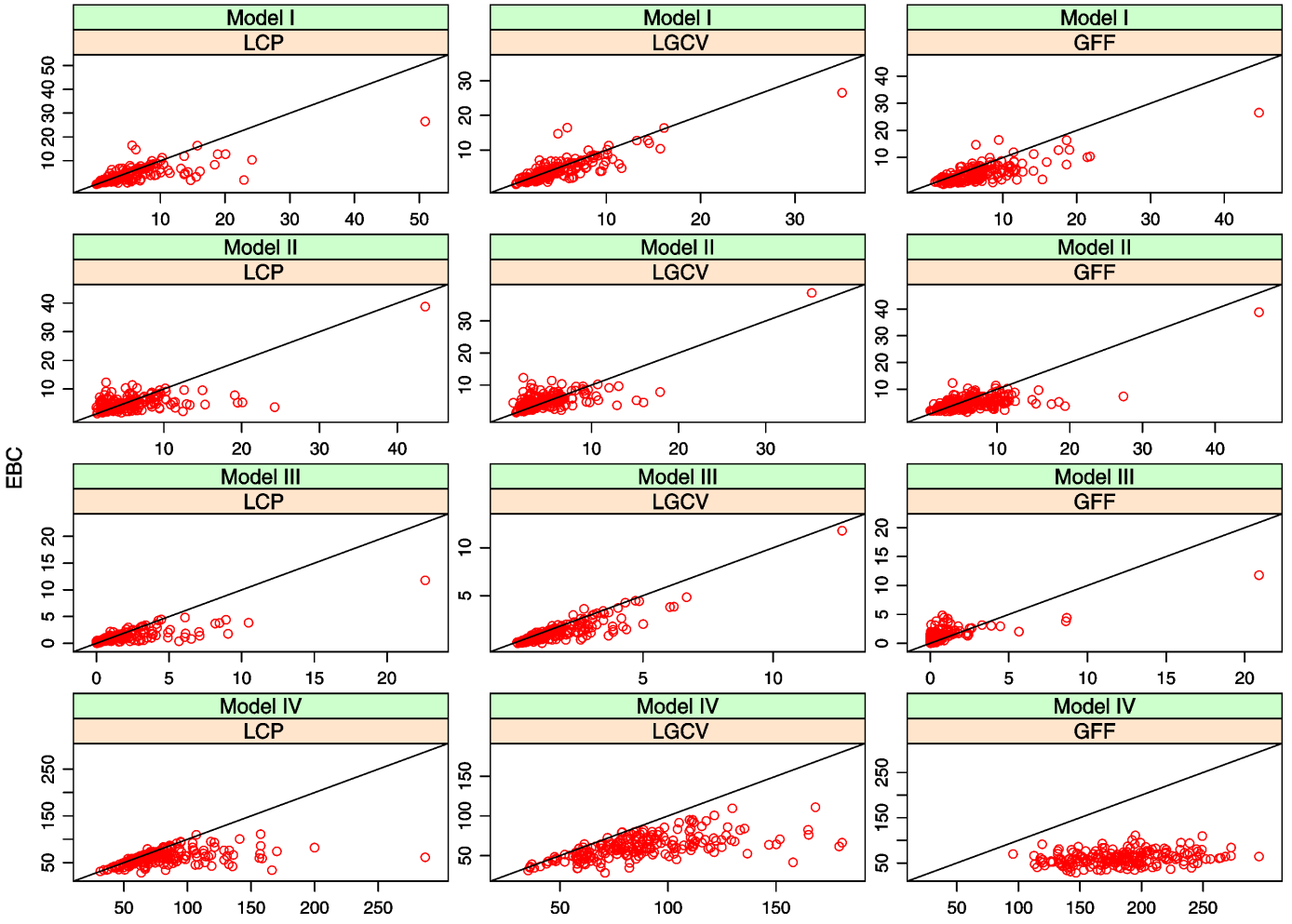

Figure 1. Pairwise Prediction Accuracy Comparison Between EBC and Other Methods for Example 1.

for models I–IV in Figure 1, and also performed paired t-tests. Table 4 reports the t statistics and p values.

Model I has a signal-to-noise ratio of approximately 5.7. The first row of Figure 1 gives the pairwise comparison between EBC and the other three methods based on the model errors for the 200 simulated datasets. We can see that EBC performs the best for this model. This is further confirmed by the paired t-tests reported in Table 4, which show that EBC yields significantly smaller model errors than the other methods.

Model II has a lower signal-to-noise ratio, approximately 1.8. As demonstrated by Table 3 and the second row of Figure 1, EBC achieves good prediction accuracy with small model size. The paired t-test results also indicate that EBC performs significantly better than LCP and GFF and similar to LGCV.

Model III represents a setup well suited for stepwise subset selection with a signal-to-noise ratio of about 7. In this case GFF, which uses a stepwise procedure, performs the best, followed by EBC.

Model IV is a bigger model, with a signal-to-noise ratio of about 9. EBC performs the best. GFF selected models with too few predictors, and consequently has a larger bias than the other methods.

In summary, LGCV tends to select models with large sizes, and its prediction performance is better in the situation where most predictors are in the true model. Because the empirical Bayes approach proposed by George and Foster (2000) can be implemented only through a stepwise greedy search, it inherits both the advantages and disadvantages of the greedy search methods. Because the algorithm is myopic, as was noted in earlier studies (Chen, Donoho, and Sauders 1999), it might work perfectly if the size of the true model is small, but in other cases it might make suboptimal choices in the first several iterations

Table 4. Paired t-Test Comparing EBC With Other Methods

	Model I		*Model II*		*Model III*		*Model IV*	
	t value	*p value*	*t value*	*p value*	*t value*	*p value*	*t value*	*p value*
EBC vs. LCP	−4.7133	0	−2.6331	.0091	−5.5673	0	−9.3160	0
EBC vs. LGCV	−3.8277	.0002	1.0483	.2958	−11.2438	0	−16.4213	0
EBC vs. GFF	−11.0830	0	−6.5922	0	6.6381	0	−46.7975	0

and end up spending most of its time correcting the mistakes made in the first few terms. In general, EBC compares favorably with other methods.

The simulation also indicates that EBC and LCP enjoy favorable computation speed. LGCV is slower, mostly because of the evaluation of the trace of the information matrix.

We ran another set of simulations that are similar to those of Breiman (1992).

Example 2. A total of 40 predictors are generated from a multivariate normal distribution with $E(x_i x_j) = .7^{|i-j|}$. The regression noise follows a standard normal distribution. Given a positive integer h, we first generate regression coefficients $\beta^*_{10+i} = \beta^*_{20+i} = \beta^*_{30+i} = (h - |i|)^2$ for integer i with $|i| < h$. The other components of β^* are set to 0. Then the coefficients are multiplied by a constant so that the theoretical $R^2 = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/(\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + n) = .75$, where n is the sample size. The simulation was conducted for each combination of three sample sizes $n = 60, 160, 600$ and five different values of $h = 1, 2, \ldots, 5$.

Figure 2 summarizes the average model error over 200 runs for each method. As the figure suggests, EBC has a clear advantage over the other methods when the sample size is small ($n = 60$). EBC and LCP perform essentially the same when the sample size is medium ($n = 60$) or larger ($n = 600$). Forward-selection–based GFF does well when the true model is sparse ($h = 1, 2$), but suffers as the true model sizes increase.

7. REAL EXAMPLES

In this section we apply the methods from Section 6 to several real datasets to compare their prediction performance. The prostate dataset, used previously by Tibshirani (1996), consists of the medical records of 97 male patients who were about to receive a radical prostatectomy. The response variable is the level of prostate-specific antigen, and there are eight predictors. The ozone data were used by Breiman and Friedman (1985), among many others. The daily maximum 1-hour average ozone reading and eight meteorologic variables were recorded in the Los Angeles Basin for 330 days of 1976. The diabetes dataset was used by Efron et al. (2004). Here 10 baseline measurements were obtained for 442 diabetes patients to predict a quantitative measure of disease progression 1 year after baseline. The Boston housing dataset concerns housing values in the suburbs of Boston (Harrison and Rubinfeld 1978), with 13 predictors collected to predict the median value of owner-occupied homes. The TLI math dataset contains math scores and demographic data of 100 randomly selected students participating in the Texas Assessment of Academic Skills (TAAS). This

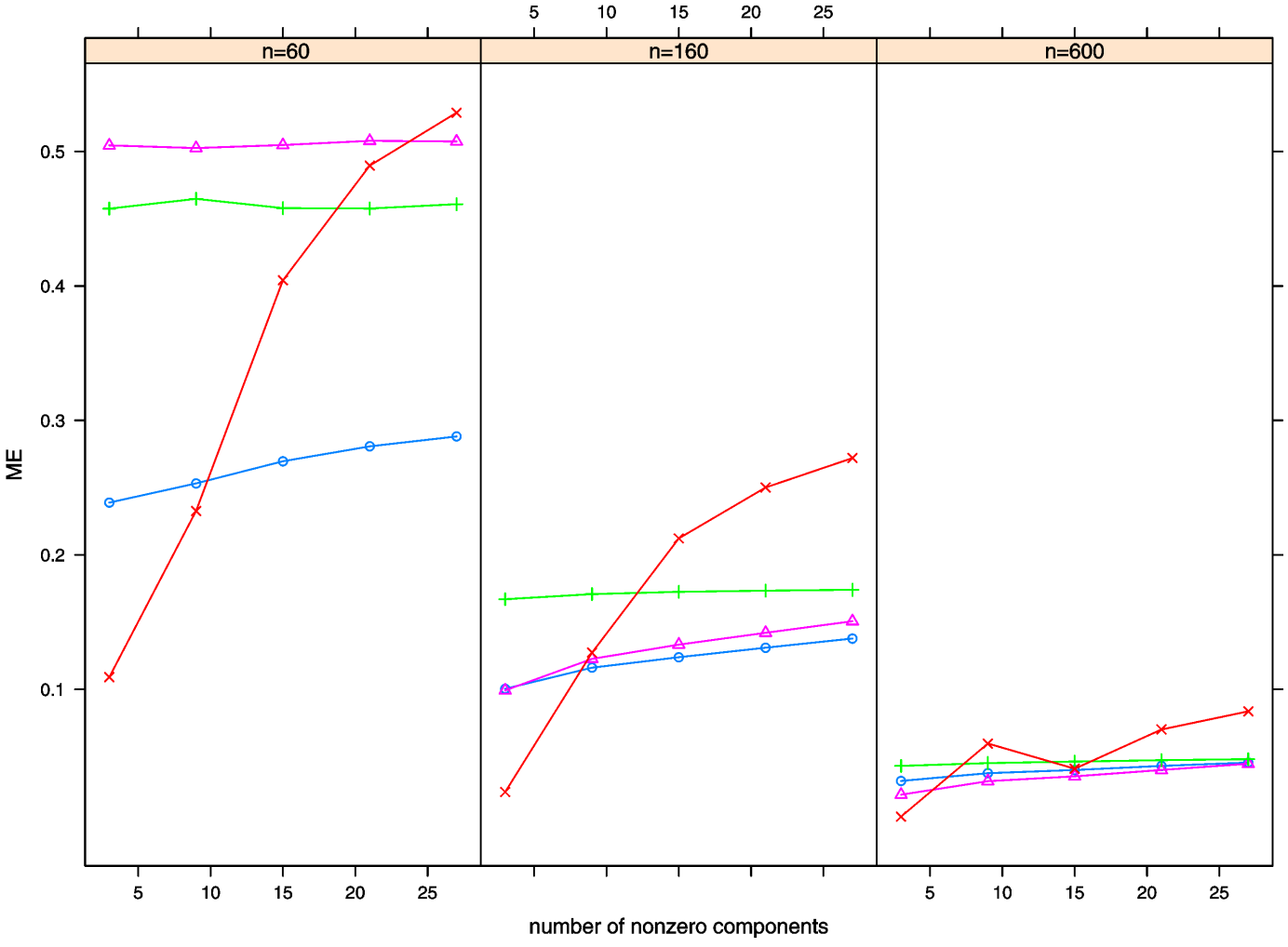

Figure 2. Prediction Accuracy for Example 2 (— ○ EBC; — △ LCP; — + LGCV; — × GFF).

dataset is available in the R package XTABLE, and more information can be obtained from the Texas Education Agency website, *http://www.tea.state.tx.us*. The dataset contains a continuous response variable, the math scores from TAAS, and four categorical explanatory variables.

Both linear model and quadratic model were considered for each of these datasets. Table 5 provides the prediction errors (PEs), average model sizes, and average CPU time consumed for different methods estimated by 10-fold cross-validation. The results in the table show that EBC and LCP enjoy a great computational advantage over LGCV and GFF. In general, EBC compares favorably with the other methods. In all examples, EBC outperforms LCP in terms of prediction accuracy.

As we pointed out earlier, a great advantage of EBC is its ability to deal with situations where $p \geq n$. To demonstrate this, we applied it to the sugar data used by Brown, Vannucci, and Fearn (2002). The goal of the experiment is to predict the composition of three sugars using near-infrared spectroscopy. There are a total of 125 training samples and 700 covariates that represent the second-difference absorbance spectra from 1,100 to 2,498 nm at 2-nm intervals. There are 21 test samples to validate the estimate. Applying EBC on the training sample results in models with 14, 20, and 16 covariates for the three sugars. The corresponding mean squared errors on the test samples are .26, .47, and .43, which are comparable with the results obtained by the much more computationally intensive approach by Brown et al. (2002). In principle, LGCV can also be applied in situations where $p \geq n$; however, it selected too many predictors in this example, and performed poorly on the test sample.

8. SUMMARY

We have developed an empirical Bayes method for variable selection and estimation in linear regression models. The method is based on a particular hierarchical Bayesian formulation, and the parameters, including the error variance in the linear model, are estimated with the data. Analytical approximations to the posterior probabilities reveal the intimate relationship between the estimator from our Bayesian formulation and the LASSO. This connection allows us to compute the Bayesian estimate with the quick LASSO algorithm. The empirical Bayes choice of the hyperparameters also provides a new way to select the tuning parameter for the LASSO algorithm.

APPENDIX A: PROOF OF THEOREM 1

Here we use the same notation as used in Section 3. Under orthogonal design, (9) can be written as

$$\begin{aligned} & C(\mathbf{Y}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\sqrt{\frac{n-1}{2\pi\sigma^2}} \right)^{|\boldsymbol{\gamma}|} \\ & \quad \times \exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{i \in \boldsymbol{\gamma}} |\beta_i|}{2\sigma^2} \right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} \\ & = C(\mathbf{Y}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\sqrt{\frac{n-1}{2\pi\sigma^2}} \right)^{|\boldsymbol{\gamma}|} \end{aligned}$$

Table 5. Comparison on Real World Examples

			EBC	*LCP*	*LGCV*	*GFF*
Prostate	Main effect	PE	.55	.57	.55	.59
($n = 97$)	($p = 8$)	Size	6.50	6.40	7.70	5.50
		Time	.06	.03	.19	.25
	Quadratic	PE	.62	.71	.79	.67
	($p = 36$)	Size	17.80	19.5	29.3	1.90
		Time	.76	.29	20.96	7.49
Diabetes	Main effect	PE	3,015.90	3,023.88	3,022.05	3,038.04
($n = 442$)	($p = 10$)	Size	7.90	7.30	9.10	8.70
		Time	.21	.10	8.10	.72
	Quadratic	PE	3,082.82	3,170.13	3,215.08	3,155.03
	($p = 64$)	Size	27.90	23.80	46.80	5.80
		Time	10.54	2.22	389.68	56.76
Ozone	Main effect	PE	21.07	21.08	21.02	20.88
($n = 330$)	($p = 8$)	Size	7.00	5.60	7.70	3.00
		Time	.09	.04	2.02	.30
	Quadratic	PE	16.24	16.67	16.29	17.2
	($p = 44$)	Size	24.50	22.40	33.40	6.40
		Time	1.73	.53	67.44	12.99
Housing	Main effect	PE	23.51	23.53	23.47	23.50
($n = 506$)	($p = 13$)	Size	11.40	11.40	13.00	13.00
		Time	.26	.11	10.23	.96
	Quadratic	PE	11.19	11.25	11.13	13.45
	($p = 103$)	Size	68.00	84.50	86.20	34.60
		Time	24.11	4.43	887.30	171.64
TLI	Main effect	PE	216.66	226.26	213.47	207.66
($n = 100$)	($p = 10$)	Size	5.10	5.40	8.70	10.00
		Time	.08	.04	1.08	.34
	Quadratic	PE	222.14	265.05	319.76	247.14
	($p = 42$)	Size	13.00	13.70	30.70	35.20
		Time	1.06	.36	5.92	5.66

$$\times \exp\left(-\frac{\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u} + \lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) d\mathbf{u}$$

$$\times \exp\left(-\frac{\|\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}\|^2 + \lambda\sum_{i\in\boldsymbol{\gamma}}|\beta_i^*|}{2\sigma^2}\right),$$

where $\tilde{\mathbf{Y}}_{\boldsymbol{\gamma}} = \mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$. Denote

$$\begin{aligned} Q &\equiv \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{\frac{n-1}{2\pi\sigma^2}}\right)^{|\boldsymbol{\gamma}|} \\ &\quad\times \exp\left(-\frac{\|\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u}\|^2 - 2\tilde{\mathbf{Y}}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{u} + \lambda\sum_{i\in\boldsymbol{\gamma}}(|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) d\mathbf{u} \\ &= \prod_{i\in\gamma}\int_{-\infty}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}} \\ &\quad\times \exp\left(-\frac{(n-1)u_i^2 - 2\tilde{\mathbf{Y}}'_{\boldsymbol{\gamma}}\mathbf{x}_i u_i + \lambda(|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) du_i \\ &\equiv \prod_{i\in\gamma} Q_i. \end{aligned} \tag{A.1}$$

a. Without loss of generality, suppose that $j \in \boldsymbol{\gamma}$ and $\beta_j^* = 0$. Let $\boldsymbol{\gamma}^*$ be the submodel of γ with the jth predictor variable excluded; then $\tilde{\mathbf{Y}}_{\boldsymbol{\gamma}} = \tilde{\mathbf{Y}}_{\boldsymbol{\gamma}^*}$ and $\beta^*_{\boldsymbol{\gamma}^*,i} = \beta^*_{\boldsymbol{\gamma},i}$, $\forall\, i \in \boldsymbol{\gamma}^*$. By (9) and (A.1), we have

$$\frac{P(\boldsymbol{\gamma}|\mathbf{Y})}{P(\boldsymbol{\gamma}^*|\mathbf{Y})} = Q_j.$$

By (15) and (16),

$$|2\tilde{\mathbf{Y}}'\mathbf{x}_j| \le \lambda.$$

This gives

$$\frac{P(\boldsymbol{\gamma}|\mathbf{Y})}{P(\boldsymbol{\gamma}^*|\mathbf{Y})} = Q_j < \int_{-\infty}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_j^2}{2\sigma^2}\right) du_j = 1.$$

The proof of part a is now completed.

b. By (9) and (A.1), we need only show that $Q_i \to 1, \forall\, i \in \boldsymbol{\gamma}$. Without loss of generality, we assume that $\operatorname{sgn}(\beta_i^*) > 0$. Similar to the derivation of (15), we have $2\tilde{\mathbf{Y}}'\mathbf{x}_i = \lambda$,

$$\begin{aligned} Q_i &= \int_{-\infty}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}} \\ &\quad\times \exp\left(-\frac{(n-1)u_i^2 - 2\tilde{\mathbf{Y}}'\mathbf{x}_i u_i + \lambda(|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) du_i \\ &= \int_{-\infty}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}} \\ &\quad\times \exp\left(-\frac{(n-1)u_i^2 + \lambda(|\beta_i^* + u_i| - \beta_i^* - u_i)}{2\sigma^2}\right) du_i \\ &= \int_{-\beta_i^*}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_i^2}{2\sigma^2}\right) du_i \\ &\quad+ \int_{-\infty}^{-\beta_i^*}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_i^2 - 2\lambda(\beta_i^* + u_i)}{2\sigma^2}\right) du_i \\ &= \Phi\left(\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) \\ &\quad+ \exp\left(\frac{\lambda^2/(n-1) + 2\lambda\beta_i^*}{2\sigma^2}\right)\Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right). \end{aligned}$$

By the mean value theorem, for some ξ between $-\frac{\beta_i^*+\lambda/(n-1)}{\sigma/\sqrt{n-1}}$ and $-\frac{\beta_i^*}{\sigma/\sqrt{n-1}}$,

$$\begin{aligned} &\Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right) - \Phi\left(-\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) \\ &\qquad = -\phi(\xi)\lambda/\sigma\sqrt{n-1} \to 0, \end{aligned}$$

given that $\lambda = o(n^{1/2})$. Thus

$$Q_i \ge \Phi\left(\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) + \Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right) \to 1.$$

In contrast,

$$\begin{aligned} &\int_{-\infty}^{-\beta_i^*}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_i^2 - 2\lambda(\beta_i^* + u_i)}{2\sigma^2}\right) du_i \\ &\qquad \le \int_{-\infty}^{-\beta_i^*}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_i^2}{2\sigma^2}\right) du_i. \end{aligned}$$

Therefore,

$$Q_i \le \int_{-\infty}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_i^2}{2\sigma^2}\right) du_i = 1.$$

Thus $Q_i \to 1$ as $n \to \infty$.

APPENDIX B: PROOF OF PROPOSITION 1

The proposition follows from two observations on h:

1. h is an decreasing function of $\boldsymbol{\gamma}$. More specifically, if $\boldsymbol{\gamma}_1$ is a submodel of $\boldsymbol{\gamma}_2$, then $h(\boldsymbol{\gamma}_1) \ge h(\boldsymbol{\gamma}_2)$.
2. $h(\widehat{\boldsymbol{\gamma}}) = h(\mathbf{1})$, where $\mathbf{1} = (1, \ldots, 1)$ represents the full model.

Combining 1 and 2, we see that for any regular model $\boldsymbol{\gamma}$,

$$h(\widehat{\boldsymbol{\gamma}}) = h(\mathbf{1}) \le h(\boldsymbol{\gamma}).$$

Now the proof is completed by the fact that $\widehat{\boldsymbol{\gamma}}$ is a regular model.

[Received May 2004. Revised February 2005.]

REFERENCES

Andrews, D. F., and Mallows, C. L. (1974), "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society*, Ser. B, 36, 99–102.

Breiman, L. (1992), "Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

——— (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598.

Brown, P. J., Vannucci, M., and Fearn, T. (2002), "Bayesian Model Averaging With Selection of Regressors," *Journal of the Royal Statistical Society*, Ser. B, 64, 519–536.

Chen, S. S., Donoho, D. L., and Sauders, M. A. (1999), "Atomic Decomposition by Basis Pursuit," *SIAM Journal of Scientific Computing*, 20, 33–61.

Chipman, H., George, E. I., and McCulloch, R. E. (2001), "The Practical Implementation of Bayesian Model Selection" (with discussion), *IMS Lecture Notes Monograph Series 38, Model Selection*, 65–134.

Efron, B., Johnston, I., Hastie, T., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

George, E. I. (2000), "The Variable Selection Problem," *Journal of the American Statistical Association*, 95, 1304–1308.

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics & Management*, 5, 81–102.

Johnstone, I. M., and Silverman, B. W. (2005), "Empirical Bayes Selection of Wavelet Thresholds," *The Annals of Statistics*, 33, 1700–1752.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1996), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Shen, X., and Ye, J. (2002), "Adaptive Model Selection," *Journal of the American Statistical Association*, 97, 210–221.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.