

Efficient Empirical Bayes Variable Selection and Estimation in Linear Models¹

Ming Yuan and Yi Lin

(February 23, 2005)

Abstract

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 only be implemented 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 when compared with other variable selection and estimation methods.

Key Words: Model selection; LASSO; LARS algorithm; Hierarchical model; Penalized least square.

¹Ming Yuan is Assistant Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA 30332-0205 (E-mail: myuan@isye.gatech.edu); and Yi Lin is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: yilin@stat.wisc.edu). Part of the work was done when Yuan was a Doctoral candidate at Department of Statistics, University of Wisconsin and supported in part by National Science Foundation grant DMS-0072292. Lin's research was supported in part by National Science Foundation grant DMS-0134987. The authors wish to thank the editor, the associate editor and two anonymous referees for comments that greatly improved the manuscript.

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, \dots, x_p) , and

$$Y = X\beta + \epsilon, \tag{1.1}$$

where $\epsilon \sim N_n(0, \sigma^2 I)$, and $\beta = (\beta_1, \dots, \beta_p)'$. Throughout this paper, we center each input variable so that the observed mean is zero, and scale each predictor so that the sample standard deviation is one.

The underlying notion behind variable selection is that some of the predictors are redundant and therefore only an unknown subset of the β 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 , AIC, and BIC, 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 of various earlier model selection methods and put them in a unified framework. The hierarchical Bayes formulation put a prior on the model space, and then put a prior on the coefficients given the model. This is conceptually attractive. George and Foster (2000) proposed to estimate the hyperparameters of the hierarchical Bayes formulation with a marginal maximum likelihood criterion or a conditional maximum likelihood 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 has to be evaluated on each candidate model to select the best model. This is impractical for even moderate number of predictors since 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; and 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, which is defined as:

$$\hat{\beta}^{LASSO}(\lambda) = \arg \min_{\beta} \left(\|Y - X\beta\|^2 + \lambda \sum |\beta_i| \right), \quad (1.2)$$

where $\lambda > 0$ is a regularization parameter. By using the L_1 penalty, minimizing (1.2) yields a sparse estimate of β if λ is chosen appropriately. Consequently, a submodel of (1.1) which contains only the covariates corresponding to the nonzero components in $\hat{\beta}^{LASSO}(\lambda)$ is selected as the final model. The LARS algorithm computes the whole path of the LASSO with a computational load in the same magnitude of the ordinary least squares. Therefore the computation is extremely fast, and this facilitates the choice of the tuning parameter with criteria such as C_p or GCV.

In this paper 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 to choose the hyperparameters, which in turn leads to an alternative method for choosing the tuning parameter in the LASSO. Unlike earlier methods including that in 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 can potentially be used in situations when the dimension is larger than the sample size.

The rest of the paper 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 which turns out to be connected to the LASSO estimate. This connection is further justified theoretically under the condition of orthogonal design in Section 4. In Section 5, we propose a method to choose 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. A summary is given in Section 8.

2 Hierarchical Model Formulation

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

- (i) a prior probability $P(\mathcal{M})$ for each candidate model \mathcal{M} ;
- (ii) a prior $P(\theta_{\mathcal{M}}|\mathcal{M})$ for parameter $\theta_{\mathcal{M}}$ associated with model \mathcal{M} ;
- (iii) 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

$$P(\mathcal{M}|Y) = \frac{P(\mathcal{M}) \int P(Y|\mathcal{M}, \theta_{\mathcal{M}})P(\theta_{\mathcal{M}}|\mathcal{M})d\theta_{\mathcal{M}}}{\sum_{\mathcal{M}'} \int P(Y|\mathcal{M}', \theta_{\mathcal{M}'})P(\theta_{\mathcal{M}'}|\mathcal{M}')d\theta_{\mathcal{M}'}P(\mathcal{M}')}. \quad (2.1)$$

We begin by indexing each candidate model with one binary vector $\gamma = (\gamma_1, \dots, \gamma_p)'$. An element γ_i takes value 0 or 1 depending on whether the i -th predictor is excluded from the model or not. Adopting this notation, under model γ , (1.1) can be written as

$$Y|\gamma, \beta \sim N\left(X_{\gamma}\beta_{\gamma}, \sigma^2 I_{(n)}\right), \quad (2.2)$$

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

Now we specify the priors for β and γ . By the definition of γ , it is natural to force $\beta_i = 0$ if $\gamma_i = 0$. On the other hand, if $\gamma_i = 1$, we give a double exponential prior for β_i . That is,

$$\beta_i|\gamma_i = (1 - \gamma_i)\delta(0) + \gamma_i DE(0, \tau), \quad j = 1, \dots, p, \quad (2.3)$$

where $DE(0, \tau)$ has density function $\tau \exp(-\tau|x|)/2$. In contrast to the commonly used normal prior $\beta_i|\gamma_i = 1 \sim N(0, \tau^2)$, the double exponential prior can better accommodate large regression coefficients due to 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 $\beta_i|\gamma_i = 1 \sim N(0, \eta_i)$. At the second level, we assume an exponential prior for η_i '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, Johnston and Silverman (2002) argued that the double exponential prior can achieve the adaptive minimax convergence rates that are 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 is a smoother alternative to our formulation, and has been adopted by George and McCulloch (1993) among others.

For γ , a widely used prior is $P(\gamma) = q^{|\gamma|}(1 - q)^{p-|\gamma|}$ with a prespecified q . This prior assumes that each predictor enters the model independently with a prior probability q , whether the predictors are correlated or not. 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 the presence of highly correlated predictors is to be avoided simply because those predictors are providing similar information on the response. To achieve this, we propose the following prior for γ

$$P(\gamma) \propto q^{|\gamma|}(1 - q)^{p-|\gamma|} \sqrt{\det(X'_\gamma X_\gamma)} \quad (2.4)$$

where $\det(X'_\gamma X_\gamma) = 1$ if $|\gamma| = 0$. Note that when the correlation between two covariates goes to 1, the prior described in (2.4) 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 $\gamma_j = 1$:

$$\frac{P(\gamma_j = 1 | \gamma^{[-j]})}{P(\gamma_j = 0 | \gamma^{[-j]})} = \frac{q}{1 - q} \sqrt{\frac{\det(X'_{\gamma^{[-j]}, \gamma_j=1} X_{\gamma^{[-j]}, \gamma_j=1})}{\det(X'_{\gamma^{[-j]}, \gamma_j=0} X_{\gamma^{[-j]}, \gamma_j=0})}}, \quad (2.5)$$

where superscript $[-j]$ indicates that the j th component is removed. If X_j is highly correlated with the current covariates, $X_{\gamma^{[-j]}, \gamma_j=0}$, the second factor of the right hand side of (2.5) will be small. Therefore, it is more likely that X_j will be removed from the full model. This is desirable since X_j does not contain much “additional” information.

Our Bayesian formulation consists of (2.2), (2.3), and (2.4). Three parameters need to be specified for this formulation, namely q , τ and σ^2 . From a hierarchical Bayesian point of view, 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 will be introduced in Section 5.

With our formulation, the joint distribution $P(\gamma, \beta_\gamma, Y)$ is

$$P(\gamma, \beta_\gamma, Y) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \frac{\sqrt{\det(X'_\gamma X_\gamma)}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|}{2\sigma^2}\right) (1-q)^p w^{|\gamma|}$$

Therefore,

$$P(\gamma|Y) = C(Y) w^{|\gamma|} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X'_\gamma X_\gamma)}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|}{2\sigma^2}\right) d\beta_\gamma, \quad (2.6)$$

where

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

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

3 Posterior Analysis

The major difficulty of the posterior inference for our Bayesian model comes from the high dimensional integration in (2.6). Since 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 which do uniformly well for all candidate models. Our proposal here is 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

$$\beta_\gamma^* = \arg \min_{\beta_\gamma} \left(\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i| \right).$$

Denote $\beta_\gamma = \beta_\gamma^* + u$. We can rewrite (2.6) as

$$P(\gamma|Y) = C(Y) w^{|\gamma|} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X'_\gamma X_\gamma)}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|}{2\sigma^2}\right) d\beta_\gamma$$

$$\begin{aligned}
&= C(Y)w^{|\gamma|} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X'_{\gamma}X_{\gamma})}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \\
&\quad \times \exp\left(-\frac{\|X_{\gamma}u\|^2 - 2\tilde{Y}'_{\gamma}X_{\gamma}u + \lambda \sum_{i \in \gamma} (|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) du \\
&\quad \times \exp\left(-\frac{\min_{\beta_{\gamma}} (\|Y - X_{\gamma}\beta_{\gamma}\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|)}{2\sigma^2}\right)
\end{aligned} \tag{3.1}$$

where $\tilde{Y}_{\gamma} = Y - X_{\gamma}\beta_{\gamma}^*$ and hereafter, we will omit the subscript γ if no confusion occurs. Our main task would be the evaluation of

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X'_{\gamma}X_{\gamma})}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|X_{\gamma}u\|^2 - 2\tilde{Y}'X_{\gamma}u + \lambda \sum_{i \in \gamma} (|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) du \tag{3.2}$$

Define

$$f(u) \equiv \frac{\|X_{\gamma}u\|^2 - 2\tilde{Y}'X_{\gamma}u + \lambda \sum_{i \in \gamma} (|\beta_i^* + u_i| - |\beta_i^*|)}{\sigma^2}. \tag{3.3}$$

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

Now we consider the following two types of models separately.

Definition 3.1 For a dataset (X, Y) and a given regularization parameter λ

- (i) a model γ is called regular if and only if β_{γ}^* does not contain zeroes or $|\gamma| = 0$;
- (ii) a model γ is called nonregular if β_{γ}^* contains at least one zero component.

3.1 Regular Models

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

$$\left. \frac{\partial^2 f(u)}{\partial u \partial u^T} \right|_{u=u^*} = \frac{1}{\sigma^2} X'_{\gamma} X_{\gamma}. \tag{3.4}$$

Applying the Laplace approximation to (3.2), we get, for sample size n large enough,

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X'_{\gamma}X_{\gamma})}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|X_{\gamma}u\|^2 - 2\tilde{Y}'X_{\gamma}u + \lambda \sum_{i \in \gamma} (|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) du \approx 1. \tag{3.5}$$

It is worth pointing out that we implicitly assumed the nonsingularity of $X'_\gamma X_\gamma$ when using the Laplace approximation. This, however, should not be a loss of generality. The models that do not satisfy this condition are not of interest from a variable selection point of view and have been assigned prior probability zero (see (2.4)). Therefore, the posterior probability for these models are always zero.

Combining (3.1) and (3.5), for sample size n large enough, we have

$$P(\gamma|Y) \approx C(Y)w^{|\gamma|} \exp\left(-\frac{\min_{\beta_\gamma} \left(\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|\right)}{2\sigma^2}\right). \quad (3.6)$$

Although (3.6) 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 independently generated from $N(0, 2^2)$ and the regression noise follows $N(0, 3^2)$. The design matrix is generated by orthogonalizing independent standard normal variables. The left hand side of (3.5) represents the ratio of the exact posterior probability to the corresponding approximation given by (3.6) for each regular model given a $\lambda > 0$. It can be factorized into univariate integrals and therefore readily computable under orthogonal design. For every $\lambda > 0$, we computed this ratio for every regular model and recorded the ratio which 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%. The experiment was then repeated 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 (3.6) is quite accurate for sample sizes as small as 20.

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
n=20	6.05%	9.37%	11.02%	11.42%	12.66%	23.89%
n=50	3.85%	5.82%	6.76%	7.00%	7.72%	14.25%
n=100	2.68%	4.04%	4.70%	4.83%	5.44%	9.34%

Table 1: Maximum Approximation Error of (3.6) for Different Sample Sizes

3.2 Nonregular Models

Although (3.6) provides a computationally efficient approximation to $P(\gamma|Y)$ for regular models, it does not apply to nonregular models since for these models $f(u)$ is not differentiable at $u = u^*$. However, we show in the following that in our model selection procedure, we can concentrate on the regular models.

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

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

which leads to

$$2\tilde{Y}'X_i = \lambda \text{sign}(\beta_{\gamma,i}^*), \quad \text{if } i \leq s. \quad (3.7)$$

On the other hand, for $s < i \leq |\gamma|$, the i th component of β_γ^* is zero, and we have

$$\left. \frac{\partial f}{\partial u_i} \right|_{u_i=0^+; u_j=0, \forall j \neq i} \geq 0, \quad \left. \frac{\partial f}{\partial u_i} \right|_{u_i=0^-; u_j=0, \forall j \neq i} \leq 0.$$

This implies that

$$|2\tilde{Y}'X_i| \leq \lambda, \quad \text{and} \quad \beta_{\gamma,i}^* = 0, \quad \text{if } s < i \leq |\gamma|. \quad (3.8)$$

By (3.7) and (3.8), from simple calculations we get,

$$\begin{aligned} & \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X_\gamma'X_\gamma)}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|X_\gamma u\|^2 - 2\tilde{Y}'X_\gamma u + \lambda \sum_{i \in \gamma} (|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2}\right) du \\ & < \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sqrt{\det(X_\gamma'X_\gamma)}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \exp\left(-\frac{\|X_\gamma u\|^2}{2\sigma^2}\right) du = 1. \end{aligned}$$

Thus,

$$P(\gamma|Y) < C(Y)w^{|\gamma|} \exp\left(-\frac{\min_{\beta_\gamma} (\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|)}{2\sigma^2}\right). \quad (3.9)$$

Now because $\tilde{Y}_\gamma = \tilde{Y}_{\gamma^*}$ and $\beta_{\gamma,i}^* = \beta_{\gamma^*,i}^*$ for any $i \leq s$, applying (3.6) to the regular model γ^* and (3.9) to the nonregular model γ , we conclude that asymptotically,

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

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

4 Connection between the LASSO and the Bayesian framework

Summarizing the above analysis, we find that if w is set to 1, then

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

$$h(\gamma) = \min_{\beta_\gamma} \left(\|Y - X_\gamma \beta_\gamma\|^2 + \lambda \sum_{i \in \gamma} |\beta_i| \right)$$

In general, these conclusions are good approximations; for the special case of orthogonal design matrix X , they can be proved rigorously:

Theorem 4.1 *Suppose that $w = 1$. Under orthogonal design, i.e. $X'X = (n-1)I_p$*

- (i) *If model γ is nonregular, then there exists a γ^* such that $P(\gamma|Y) < P(\gamma^*|Y)$*
- (ii) *Suppose that $\lambda = o(\sqrt{n})$. If model γ is regular, then as $n \rightarrow \infty$,*

$$P(\gamma|Y) \sim C(Y) \exp \left(-\frac{\|Y - X_\gamma \beta_\gamma^*\|^2 + \lambda \sum_{i \in \gamma} |\beta_i^*|}{2\sigma^2} \right).$$

We note the $(n-1)$ factor in the condition $X'X = (n-1)I_p$ is just to conform to our convention we made at the beginning of the paper that the data be scaled to have sample standard deviation one.

By (i) and (ii), we can now focus on searching for the regular model γ with the smallest $h(\gamma)$. A straightforward search involves going through each of the large number of candidate

models to identify regular models and to minimize h over all the regular models. This would be computationally very demanding. Fortunately, such a search is not necessary, and the following proposition provides us with the key to a simple and explicit recipe for finding a regular model that minimizes $h(\gamma)$. The proof is relegated to the appendix.

Proposition 4.1 *Let $\hat{\beta} = \arg \min_{\beta} (\|Y - X\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|)$, and let model $\hat{\gamma}$ be such that $\hat{\gamma}_i = I(\hat{\beta}_i \neq 0)$, where $I(\cdot)$ is the indicator function. Then $\hat{\gamma}$ is the regular model that minimizes $h(\gamma)$.*

Interestingly, $\hat{\gamma}$ is exactly the same model selected by the LASSO. In other words, the model selected by the LASSO has the highest posterior probability under our Bayesian model (2.4) 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 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, we can use the LASSO 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 in George and Foster (2000) which can only be implemented 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. Tibshirani (1996) mentioned that the LASSO has another Bayesian interpretation with independent double exponential prior on each regression coefficient. Tibshirani's formulation is somehow less natural as a Bayesian variable selection procedure since it puts prior probability one on the full model. Consequently, the corresponding posterior probability for the full model will also be one even if the posterior modal estimates of some regression coefficients are zero.

While setting $w = 1$ substantially eases the computational burden, it may potentially incur loss of efficiency in terms of prediction accuracy. However, we found that this potential loss of efficiency is usually small. To illustrate, we did a small experiment where $y_i = \beta_i + \epsilon_i$, $i = 1, \dots, 500$. 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 fifty β'_i s are generated from $N(0, 1)$ and the rest are set to be 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 one hundred 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 $\hat{\beta}_{w,\lambda}$. Consequently, we can compute $\hat{\beta}_{opt} = \arg \min_{w,\lambda} \|\hat{\beta}_{w,\lambda} - \beta\|^2$ and $\hat{\beta}_{opt,w=1} = \arg \min_{\lambda} \|\hat{\beta}_{1,\lambda} - \beta\|^2$. The summary statistics of the relative estimation efficiency $\|\hat{\beta}_{opt} - \beta\|^2 / \|\hat{\beta}_{opt,w=1} - \beta\|^2$ over 100 runs are reported in Table 2.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
66.67%	92.66%	97.32%	95.01%	99.36%	100.00%

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

As Table 2 clearly indicates, the loss of efficiency caused by forcing $w = 1$ is small. Therefore it is reasonable to set w to one, given the great computational advantage 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. Tibshirani (1996) proposed a GCV score to select λ . In the following, we adopt an empirical Bayesian approach for selecting both σ^2 and λ .

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

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

This can be implemented when the number of variables is small. However, 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 that is related to the conditional maximum “likelihood” approach proposed in George and Foster

(2000). The conditional density of Y given a model γ is

$$f(Y|\gamma, \sigma^2, \lambda) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{\|Y - X_\gamma \beta_\gamma\|^2}{2\sigma^2}\right) \\ \times \left(\frac{\lambda}{4\sigma^2} \right)^{|\gamma|} \exp\left(-\frac{\lambda \sum_{i \in \gamma} |\beta_i|}{2\sigma^2}\right) d\beta_\gamma.$$

For a given λ , denote the selected model by $\hat{\gamma}_\lambda$. We choose σ^2 and λ as the maximizer of $f(Y|\hat{\gamma}_\lambda, \sigma^2, \lambda)$. Strictly speaking, $f(Y|\hat{\gamma}_\lambda, \sigma^2, \lambda)$ is not a likelihood nor a conditional likelihood, and we denote the resulting criterion by CML only because of its similarity to the approach in George and Foster (2000). Since $\hat{\gamma}_\lambda$ is regular, we can use (3.1) and (3.5) to approximate $f(Y|\hat{\gamma}_\lambda, \sigma^2, \lambda)$.

$$f(Y|\hat{\gamma}_\lambda, \sigma^2, \lambda) \approx \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^{n-|\hat{\gamma}_\lambda|} \left(\frac{\lambda}{4\sigma^2} \right)^{|\hat{\gamma}_\lambda|} (\det(X'_{\hat{\gamma}_\lambda} X_{\hat{\gamma}_\lambda}))^{-1/2} \times \\ \times \exp\left(\frac{-\min_{\beta} (\|Y - X_{\hat{\gamma}_\lambda} \beta_{\hat{\gamma}_\lambda}\|^2 + \lambda \sum_{i \in \hat{\gamma}_\lambda} |\beta_i|)}{2\sigma^2}\right) \\ = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^{n-|\hat{\gamma}_\lambda|} \left(\frac{\lambda}{4\sigma^2} \right)^{|\hat{\gamma}_\lambda|} (\det(X'_{\hat{\gamma}_\lambda} X_{\hat{\gamma}_\lambda}))^{-1/2} \times \\ \times \exp\left(-\frac{\min_{\beta} (\|Y - X\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|)}{2\sigma^2}\right)$$

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

$$\text{CML}(\lambda) \equiv (n + |\hat{\gamma}_\lambda|) \left[\ln \left(\frac{\min_{\beta} (\|Y - X\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|)}{n + |\hat{\gamma}_\lambda|} \right) + 1 \right] \\ + \ln (\det(X'_{\hat{\gamma}_\lambda} X_{\hat{\gamma}_\lambda})) - 2|\hat{\gamma}_\lambda| \ln(\sqrt{2\pi}\lambda/4),$$

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

Remark Both C_p used in 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 therefore can only be applied in the situation $p < n$. Our proposal here avoids this problem.

Remark 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, since the approximation only applies to regular models. However, it is tempting to maximize

$f(Y|\hat{\beta}_\lambda, \hat{\gamma}_\lambda, \sigma^2, \lambda)$ instead. Unfortunately, this leads to a criterion

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

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

6 Simulations

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

- (i) (EBC) Our approximate empirical Bayes estimate with hyper-parameters selected by CML;
- (ii) (LCP) The LASSO with λ selected by C_p ;
- (iii) (LGCV) The LASSO with λ selected by GCV;
- (iv) (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 to choose the hyper-parameters 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 750M computer. All simulations were conducted using R. The path of the LASSO estimate was computed using the package LARS in R. The model error of an estimate $\hat{\beta}$ is given by

$$ME(\hat{\beta}) = (\hat{\beta} - \beta)' V (\hat{\beta} - \beta),$$

where $V = E(X'X)$ is the population covariance matrix of X . The models in our first simulation example have also been used in Tibshirani (1996).

Example 6.1 *Consider the following four models*

- I. $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\sigma = 3$. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$.
- II. Same as (I) except that $\beta_j = 0.85$ for all j .

III. Same set-up as before, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ and $\sigma = 2$.

IV. Forty correlated predictors are considered. $x_{ij} = z_{ij} + w_i$, where z_{ij}, 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 the first three models, two hundred datasets with sample size 20 were generated. For the fourth model, two hundred datasets with sample size 100 were generated.

	EBC		LCP		LGCV		GFF	
Model I								
ME	3.99	(0.24)	5.19	(0.37)	4.52	(0.26)	6.37	(0.34)
Size	5.14	(0.08)	5.29	(0.11)	7.37	(0.05)	5.66	(0.21)
Time (sec)	0.11	(0.00)	0.05	(0.00)	0.23	(0.00)	0.27	(0.00)
Model II								
ME	4.95	(0.23)	5.60	(0.32)	4.76	(0.25)	6.55	(0.33)
Size	5.68	(0.08)	5.70	(0.10)	7.22	(0.06)	6.80	(0.18)
Time (sec)	0.11	(0.00)	0.05	(0.00)	0.23	(0.00)	0.27	(0.00)
Model III								
ME	1.19	(0.09)	1.81	(0.17)	1.70	(0.10)	0.64	(0.13)
Size	4.23	(0.09)	4.06	(0.16)	7.26	(0.05)	1.32	(0.09)
Time (sec)	0.11	(0.00)	0.05	(0.00)	0.23	(0.00)	0.29	(0.00)
Model IV								
ME	61.18	(1.05)	80.22	(2.26)	87.18	(1.98)	183.47	(2.66)
Size	25.45	(0.16)	27.26	(0.32)	33.43	(0.22)	6.89	(0.09)
Time (sec)	0.87	(0.01)	0.30	(0.00)	5.01	(0.03)	8.96	(0.01)

Table 3: Comparisons on the Simulated Datasets

Table 3 gives the means and standard errors (in the parentheses) over the 200 simulated datasets. We see from the tables that EBC tends to select models with relatively smaller sizes than the other methods. To see whether the choice of sparse models comes with sacrificing the prediction accuracy, we provide a pairwise prediction accuracy comparison between EBC

and the other methods for Model I-IV in Figure 1, and also perform paired t-tests. Table 4 reports the t-statistics and p-values.

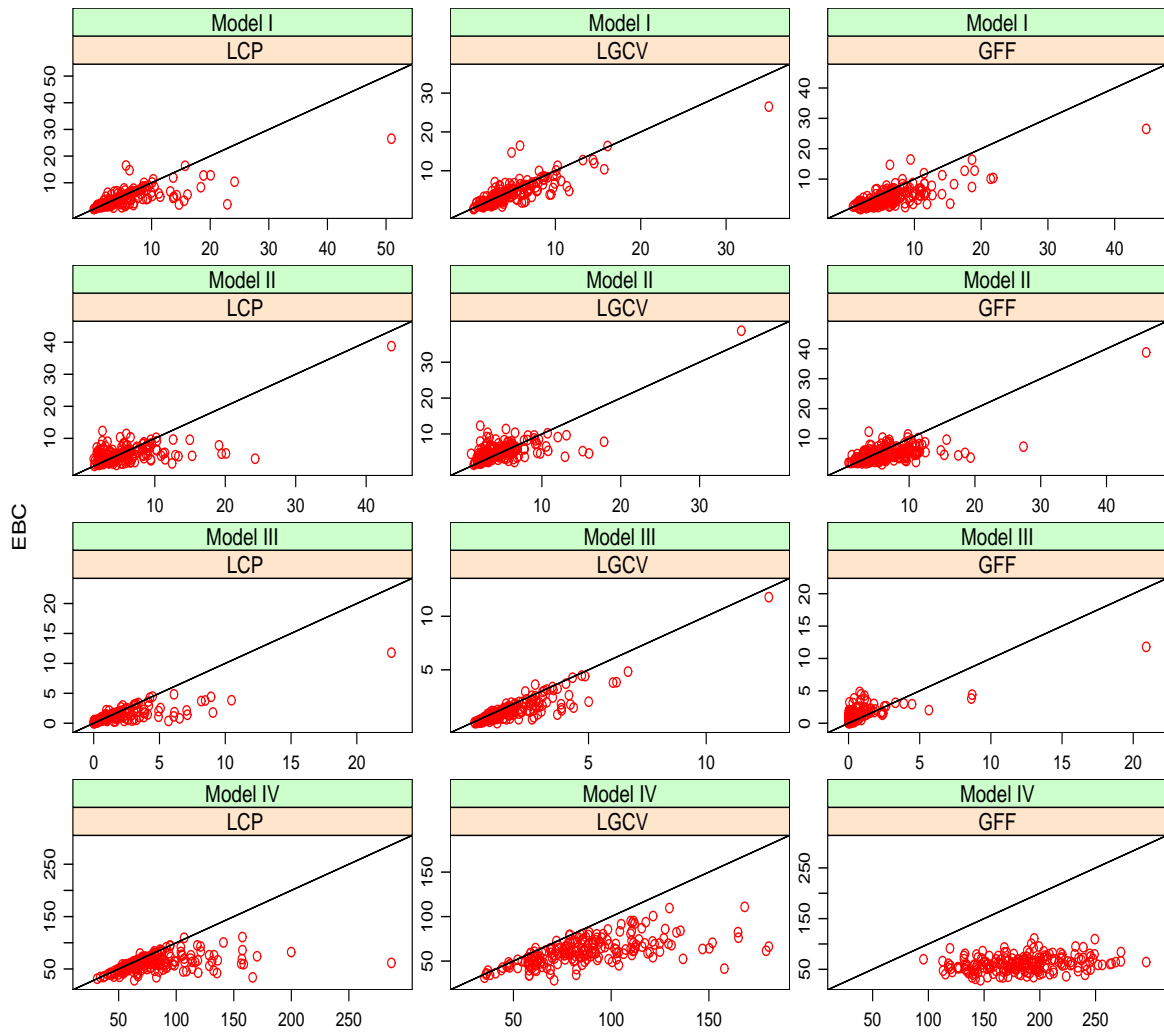


Figure 1: Pairwise Prediction Accuracy Comparison between EBC and Other Methods for Example 6.1

Model I has a signal to noise ratio 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 shows that EBC yields significantly smaller model errors than the 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	0.0000	-2.6331	0.0091	-5.5673	0.0000	-9.3160	0.0000
EBC vs LGCV	-3.8277	0.0002	1.0483	0.2958	-11.2438	0.0000	-16.4213	0.0000
EBC vs GFF	-11.0830	0.0000	-6.5922	0.0000	6.6381	0.0000	-46.7975	0.0000

Table 4: Paired t-test Comparing EBC with Other Methods

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

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

Model IV is a bigger model with signal to noise ratio 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. Since the empirical Bayes approach proposed by George and Foster (2000) can only be implemented through a stepwise greedy search, it inherits both the advantages and disadvantages of the greedy search methods. Because the algorithm is myopic, as noted in earlier studies (Chen, Donoho and Saunders, 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 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 in Breiman (1992).

Example 6.2 *Forty predictors are generated from a multivariate normal distribution with $E(x_i x_j) = 0.7^{|i-j|}$. 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 $R^2 = \beta' X' X \beta / (\beta' X' X \beta + n) = 0.75$, 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 $h = 1, 2, \dots, 5$.

Figure 2 summarizes the average model error over 200 runs for each method. As Figure 2 suggests, EBC shows 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.

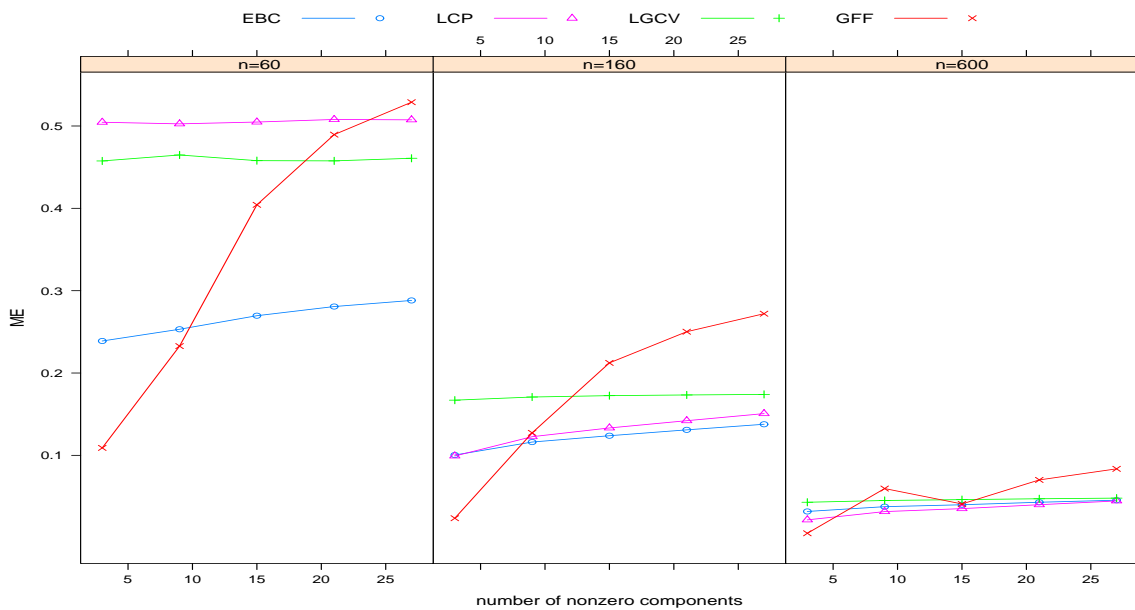


Figure 2: Prediction Accuracy for Example 6.2

7 Real Examples

In this section, we apply the methods from Section 6 to several real datasets to compare their prediction performance. The prostate data, previously used in 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 8 predictors. The

ozone data was used in Breiman and Friedman (1985) among many other references. The daily maximum one-hour-average ozone reading and 8 meteorologic variables were recorded in the Los Angeles Basin for 330 days of 1976. The diabetes dataset was used in Efron et. al. (2004). Ten baseline measurements were obtained for 442 diabetes patients in order to predict a quantitative measure of disease progression one year after baseline. The boston housing data concerns housing values in the suburbs of Boston (Harrison and Rubinfeld, 1978). Thirteen predictors were collected in order to predict 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 dataset is available in the XTABLE package of R and more information can be obtained from the website of Texas Education Agency at <http://www.tea.state.tx.us>. The dataset contains a continuous response variable, the math scores from TAAS and 4 categorical explanatory variables.

For each of these datasets, both linear model and quadratic model were considered. Table 5 provides the prediction errors (PE), 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 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 before, a great advantage possessed by 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 which represent the second difference absorbance spectra from 1100 to 2498 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 respectively for the three sugars. The corresponding mean squared errors on the test samples are 0.26, 0.47 and 0.43 respectively, which are comparable with the results obtained by a much more computationally intensive approach in Brown, Vannucci and Fearn (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.

			EBC	LCP	LGCV	GFF
Prostate (n=97)	Main Effect (p=8)	PE	0.55	0.57	0.55	0.59
		Size	6.50	6.40	7.70	5.50
		Time	0.06	0.03	0.19	0.25
	Quadratic (p=36)	PE	0.62	0.71	0.79	0.67
		Size	17.80	19.5	29.3	1.90
		Time	0.76	0.29	20.96	7.49
Diabetes (n=442)	Main Effect (p=10)	PE	3015.90	3023.88	3022.05	3038.04
		Size	7.90	7.30	9.10	8.70
		Time	0.21	0.10	8.10	0.72
	Quadratic (p=64)	PE	3082.82	3170.13	3215.08	3155.03
		Size	27.90	23.80	46.80	5.80
		Time	10.54	2.22	389.68	56.76
Ozone (n=330)	Main Effect (p=8)	PE	21.07	21.08	21.02	20.88
		Size	7.00	5.60	7.70	3.00
		Time	0.09	0.04	2.02	0.30
	Quadratic (p=44)	PE	16.24	16.67	16.29	17.2
		Size	24.50	22.40	33.40	6.40
		Time	1.73	0.53	67.44	12.99
Housing (n=506)	Main Effect (p=13)	PE	23.51	23.53	23.47	23.50
		Size	11.40	11.40	13.00	13.00
		Time	0.26	0.11	10.23	0.96
	Quadratic (p=103)	PE	11.19	11.25	11.13	13.45
		Size	68.00	84.50	86.20	34.60
		Time	24.11	4.43	887.30	171.64
TLI (n=100)	Main Effect (p=10)	PE	216.66	226.26	213.47	207.66
		Size	5.10	5.40	8.70	10.00
		Time	0.08	0.04	1.08	0.34
	Quadratic (p=42)	PE	222.14	265.05	319.76	247.14
		Size	13.00	13.70	30.70	35.20
		Time	1.06	0.36	5.92	5.66

Table 5: Comparison on Real World Examples

8 Summary

We 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.

References

- [1] Andrews, D. F. and Mallows, C. L. (1974), Scale mixtures of normal distributions, *J. R. Statist. Soc. B*, **36**, 99-102.
- [2] Breiman, L. (1992), Little Bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error, *J. Amer. Statist. Assoc.*, **87**, 738-754.
- [3] Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373-384.
- [4] Breiman, L. and Friedman, J. H. (1985), Estimating optimal transformations for multiple regression and correlation, *J. Amer. Statist. Assoc.*, **80**, 580-598.
- [5] Brown, P. J., Vannucci, M. and Fearn, T. (2002), Bayesian model averaging with selection of regressors, *J. R. Statist. Soc. B*, **64**, 519-536.
- [6] Chen, S. S., Donoho, D. L. and Saunders, M. A. (1999), Atomic Decomposition by Basis Pursuit, *SIAM J. Scientific Computing*, **20**, 33-61.
- [7] Chipman, H., George, E. I. and McCulloch, R. E. (2001), The practical implementation of Bayesian model selection (with discussion), *IMS Lecture Notes Monogr. Ser.*, *38*, *Model selection*, 65-134.

- [8] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004), Least angle regression, *Ann. Statist.*, **32** 407-499.
- [9] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96** 1348-1360.
- [10] Foster, D. P. and George, E. I. (1994), The risk inflation criterion for multiple regression, *Ann. Statist.*, **22**, 1947-1975.
- [11] George, E. I. (2000), The variable selection problem, *J. Amer. Statist. Assoc.*, **95**, 1304-1308.
- [12] George, E. I. and Foster, D. P. (2000), Calibration and empirical Bayes variable selection, *Biometrika*, **87**, 731-747.
- [13] George, E. I. and McCulloch, R. E. (1993), Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, **88**, 881-889.
- [14] Harrison, D. and Rubinfeld, D.L. (1978), Hedonic prices and the demand for clean air, *J. Environ. Economics & Management*, **5**, 81-102.
- [15] Johnston, I. and Silverman, B. W. (2002), Empirical Bayes selection of wavelet thresholds.
- [16] Raftery, A. E., Madigan, D. and Hoeting, J. A. (1996), Bayesian model averaging for linear regression models, *J. Amer. Statist. Assoc.*, **92**, 179-191.
- [17] Shen, X. and Ye, J. (2002) Adaptive model selection, *J. Amer. Statist. Assoc.*, **97**, 210-221.
- [18] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, **58**, 267-288.

Appendix A – Proof of Theorem 4.1

Proof. We use the same notation as those used in Section 3. Under orthogonal design, (3.1) can be written as

$$\begin{aligned}
& C(Y) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\sqrt{\frac{n-1}{2\pi\sigma^2}} \right)^{|\gamma|} \exp \left(-\frac{\|Y - X_{\gamma}\beta_{\gamma}\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|}{2\sigma^2} \right) d\beta_{\gamma} \\
&= C(Y) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\sqrt{\frac{n-1}{2\pi\sigma^2}} \right)^{|\gamma|} \\
&\quad \times \exp \left(-\frac{\|X_{\gamma}u\|^2 - 2\tilde{Y}'_{\gamma}X_{\gamma}u + \lambda \sum_{i \in \gamma} (|\beta_i^* + u_i| - |\beta_i^*|)}{2\sigma^2} \right) du \\
&\quad \times \exp \left(-\frac{\|Y - X_{\gamma}\beta_{\gamma}^*\|^2 + \lambda \sum_{i \in \gamma} |\beta_i^*|}{2\sigma^2} \right)
\end{aligned}$$

where $\tilde{Y}_{\gamma} = Y - X_{\gamma}\beta_{\gamma}^*$. Denote

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

- (i) Without loss of generality, suppose that $j \in \gamma$ and $\beta_j^* = 0$. Let γ^* be the submodel of γ with the j -th predictor variable excluded, then $\tilde{Y}_{\gamma} = \tilde{Y}_{\gamma^*}$ and $\beta_{\gamma^*,i}^* = \beta_{\gamma,i}^*$, $\forall i \in \gamma^*$. By (3.1) and (A.1), we have,

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

By (3.7) and (3.8),

$$|2\tilde{Y}'x_j| \leq \lambda.$$

This gives

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

The proof of (i) is now completed.

(ii) By (3.1) and (A.1), we only need to show that $Q_i \rightarrow 1, \forall i \in \gamma$. Without loss of generality, we assume that $\text{sgn}(\beta_i^*) > 0$. Similar to the derivation of (3.7), we have $2\tilde{Y}'x_i = \lambda$.

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

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

$$\Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right) - \Phi\left(-\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) = -\phi(\xi)\lambda/\sigma\sqrt{n-1} \rightarrow 0$$

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

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

On the other hand,

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

Therefore,

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

Thus, $Q_i \rightarrow 1$ as $n \rightarrow \infty$. ■

Appendix B – Proof of Proposition 4.1

Proof. The proposition follows from two observations on h .

(i) h is an decreasing function of γ . More specifically, if γ_1 is a submodel of γ_2 , then

$$h(\gamma_1) \geq h(\gamma_2);$$

(ii) $h(\hat{\gamma}) = h(\mathbf{1})$, where $\mathbf{1} = (1, \dots, 1)$ stands for the full model.

Combining (i) and (ii), we see that for any regular model γ ,

$$h(\hat{\gamma}) = h(\mathbf{1}) \leq h(\gamma).$$

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