Journal of Data Science **10**(2012), 363-383

A Bimodal Spike and Slab Model for Variable Selection and Model Exploration

Tanujit Dey
College of William & Mary

Abstract: We have developed an enhanced spike and slab model for variable selection in linear regression models via restricted final prediction error (FPE) criteria; classic examples of which are AIC and BIC. Based on our proposed Bayesian hierarchical model, a Gibbs sampler is developed to sample models. The special structure of the prior enforces a unique mapping between sampling a model and calculating constrained ordinary least squares estimates for that model, which helps to formulate the restricted FPE criteria. Empirical comparisons are done to the lasso, adaptive lasso and relaxed lasso; followed by a real life data example.

Key words: FPE analysis, model exploration, rescaled spike and slab model, variable selection.

1. Introduction

We consider the problem of selecting variables in a linear regression model. To outline the problem mathematically, we consider the linear regression model

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_K x_{i,K} + \varepsilon_i = \mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where the responses $Y_1, \cdots, Y_n$ are independent with corresponding K-dimensional predictors $\mathbf{x}_1, \cdots, \mathbf{x}_n$. The $\{\varepsilon_i\}$'s are independent variables such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma^2 > 0$. The dilemma is to find the subset of nonzero covariates from $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_K)^t$. In this article it is also assumed that $\mathbf{x}_i$ are standardized so that $\sum_{i=1} x_{i,k} = 0$ and $\sum_{i=1} x_{i,k}^2 = n$ for each k.

The purpose of this article is to use final prediction error (FPE) (Akaike, 1969) criteria for variable selection via a restricted model search mechanism based on Bayesian hierarchical model.

FPE is a criteria which takes the residual sum of squares (RSS) and tacks on a penalty related to the number of variables. To properly define FPE for a subset $\alpha \subseteq \{1, \cdots, K\}$, let β_α be the components of $\beta = (\beta_1, \cdots, \beta_K)^t$ which are

indexed by the elements of α. The formal definition of FPE criteria based on the linear regression model (1) is:

$$\text{FPE}(\alpha) = n^{-1}\text{RSS}(\alpha) + \lambda_n n^{-1}\hat{\sigma}^2 K_\alpha. \tag{2}$$

In (2), $\text{RSS}(\alpha)$ is the RSS for model α when the first K_α variables are entered in the model, K_α is the size of the model α, $\hat{\sigma}^2$ is the unbiased estimator for σ^2 and λ_n is the non-zero penalty associated with the model. Model selection based on FPE chooses the model of size $K_{\hat{\alpha}}$ which minimizes (2) with respect to α. Classic examples of the FPE criteria are AIC (Akaike, 1973) and BIC (Schwarz, 1978). Note that, expression (2), setting $\lambda_n = 2$ generates the AIC criterion and the BIC criterion if $\lambda_n = \log(n)$.

In general to implement FPE variable selection approach; we need to visit all 2^K models and then apply the criteria to select the best subset of variables. Unfortunately this is not viable in high dimensional setting as it becomes computationally infeasible even for relatively small K. Instead we use a Bayesian approach based on rescaled spike and slab models (Ishwaran and Rao, 2003, 2005 a, b); a variable selection method for linear regression models, which is rooted in the spike and slab models (George and McCulloch, 1993). Using this model we design a Gibbs sampler which allow us to draw values from the Bayesian posterior. This Gibbs sampler is highly efficient and is capable of effective search across the relevant model space which results in a restricted all subset search. Using these models we can implement a restricted FPE variable selection technique.

Another goal is to explore model space based on Bayesian hierarchical models. Model space exploration is extremely important from a machine learning perspective, especially in high dimensional setting. Therefore without executing any type of variable selection technique, we are able to locate variables that are vital in explaining the data.

The restricted FPE variable selection approach is solely driven by the proposed Bayesian hierarchical models; therefore, it is vital to study the impact of this model for searching the entire model space. By model space exploration we mean the following: the Gibbs sampler samples posterior values in each draw to produce desired model for performing restricted FPE analysis. These values also drive the posterior model probability, a higher probability value results in a larger model and vice versa. If Gibbs sampler samples only smaller models or only larger models, then resulting estimator will be underestimated or overestimated respectively. To analyze this issue, we theoretically study the performance of the proposed model in exploring the model space on the basis of Bayesian model averaged (BMA) estimator. Theoretical results show that Gibbs sampler based on our proposed model; samples all possible model sizes required for exploring the model space. Graphical tools are useful especially in high dimensional setting, to

pinpoint the significance of variables in a given data set. We present a graphical tool to do so.

The article is organized as follows: Section 2 reviews the spike and slab hierarchical models and proposes a new bimodal prior in the spike and slab hierarchy. Also discussed is how this special type of hypervariance structure of the prior helps to compute FPE estimates for several models. Section 3 discusses the interplay between the posterior mean which is the BMA estimator and model space exploration. Section 4 introduces rescaled spike and slab models (termed the "bimodal spike and slab model"). The remainder of this Section talks about how this model overcomes certain limitations of earlier proposed model in Section 2. In Section 5; a simulation study compares our proposed method with other popular model selection methods from the perspective of model space exploration; another simulation study is done to study the performance of the proposed model in the context of high-dimensional sparse linear model. Section 6 has real life data analysis using the R package modelSampler, developed based on our proposed model. The article concludes in Section 7. The proof of all lemmas and theorems of this article are placed in the Appendix.

2. A New Spike and Slab Model

Spike and slab models (Ishwaran and Rao, 2005b; George and McCulloch, 1993, 1997; Mitchell and Beauchamp, 1988) are a popular methodology for selecting variables in the regression set up in (1). A spike and slab model defined in Ishwaran and Rao (2005b) is a Bayesian model specified by the following prior hierarchy:

$$\begin{aligned}(Y_i|\mathbf{x}_i,\beta,\hat{\sigma}^2) &\overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{x}_i^t\beta,\hat{\sigma}^2), \quad i=1,\cdots,n,\\ (\beta|\boldsymbol{\gamma}) &\sim \mathrm{N}(\mathbf{0},\boldsymbol{\Gamma}),\\ \boldsymbol{\gamma} &\sim \pi(d\boldsymbol{\gamma}).\end{aligned} \tag{3}$$

Here $\mathbf{0}$ is the K-dimensional zero vector, $\boldsymbol{\Gamma}$ is the $K\times K$ diagonal matrix $\mathrm{diag}(\gamma_1,\cdots,\gamma_K)$ and π is the prior measure for $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_K)^t$.

All through this article the prior π for $\boldsymbol{\gamma}$ is specified as:

$$\begin{aligned}(\gamma_k|v_0,V,w) &\overset{\text{iid}}{\sim} (1-w)\,\delta_{v_0}(\cdot)+w\,\delta_V(\cdot), \quad k=1,\cdots,K,\\ w &\sim \mathrm{Uniform}[0,1],\end{aligned} \tag{4}$$

where $\delta_v(\cdot)$ is a discrete measure concentrated at the value v. This prior for γ is in essence a two-component mixture model with one component taking on a very small value $v_0>0$, and the other component taking on a very large value $V>0$. It induces a prior for β, which is a mix of near-degenerate multinormal distributions concentrated on a submodel α. An important feature of

this two-component prior is that it has a selective shrinkage property. Selective shrinkage is a property where the posterior mean shrinks towards zero only for coefficients that are truly zero. For theoretical justification of this feature for a two-component prior see Ishwaran and Rao (2011). Another important feature of (4) is the presence of w, which we label as "complexity parameter", that controls the size of the models. Note that using an indifference prior is equivalent to choosing a degenerate prior for w at the value of $1/2$. Using a continuous prior for w, therefore allows for a greater amount of adaptiveness in estimating model size.

As stated earlier; the purpose of this article is to explore model and at the same time; perform a restricted FPE analysis for variable selection. The implementation of the FPE analysis is done using the following procedure.

Running the Gibbs sampler, we track the different α models as they are being sampled. The model α is identified by introducing the following notation: Defining

$$I_k = \begin{cases} 1, & \text{if } \gamma_k = V, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Thus, each draw for $\boldsymbol{\gamma}$ has an associated binary K-tuple $(I_1, \cdots, I_K)$. Therefore the model α is the model where $\alpha = \{k : I_k = 1\}$. This creates a unique mapping between $\boldsymbol{\gamma}$ and α. Once we have the model α, we calculate the constrained OLS estimates of that model based on the indices of α, and then calculate the RSS for the model. Computation of AIC and BIC values are carried out using (2), which allows us to study the empirical performance of methods like AIC and BIC without being restricted to the number of predictors. This is greatly beneficial since classical implementations of an all subsets search (such as leaps-and-bounds (Furnival and Wilson, 1974)) is typically restricted to a handful of predictors, such as $K = 30$. The steps for calculating the FPE criteria are as follows. If the Gibbs sampler visits a model α, we calculate the constrained OLS of model α as

$$\hat{\beta}_{+\alpha} = (\mathbf{X}_{+\alpha}^t \mathbf{X}_{+\alpha} + \mathbf{I}_{+\alpha})^{-1} \mathbf{X}_{+\alpha}^t \mathbf{Y}, \tag{6}$$

where $+\alpha$ indicates the set of non-zero elements in α and $\mathbf{I}_{+\alpha}$ is an identity matrix of order $(+\alpha \times +\alpha)$. The corresponding RSS of that model will be $\text{RSS}(\alpha) = ||\mathbf{Y} - \mathbf{X}_{+\alpha}\hat{\beta}_{+\alpha}||^2$. Once we have the $\text{RSS}(\alpha)$, we calculate the FPE defined in (2).

3. The Effect of Model Exploration on the Posterior Mean

This section shows in what ways model exploration can affect the BMA estimator, which is the posterior mean. Our discussion is based on the models (3)

and (4). To define BMA formally, let $\hat{\beta}$ be the BMA estimator of β, then

$$\hat{\beta} = \mathbb{E}(\beta|\mathbf{Y}) = \sum_{\alpha} \boldsymbol{\mu}_{\alpha} \Pr(\alpha|\mathbf{Y}), \tag{7}$$

where $\boldsymbol{\mu}_{\alpha} = \mathbb{E}(\beta|\alpha, \mathbf{Y})$, is the conditional posterior mean under model α and $\Pr(\alpha|\mathbf{Y})$ is the posterior model probability of model α.

The BMA estimators are often used in Bayesian paradigm for model selection as BMA is a stable estimator since averaging over all the models takes care of model uncertainty, therefore it has better predictive performance. From (7), we notice that the BMA is driven by the posterior model probability; a higher value of probability results in a better estimator, whereas a smaller value presents an estimator of inferior quality. We show that the conditional posterior mean is approximately equivalent to constrained OLS estimator under our proposed models. Therefore; BMA is driven by the posterior model probability and constrained OLS estimator. In (5) and (6), we show how to calculate constrained OLS estimators. Hence if the Gibbs sampler moves around a smaller model space, it results in exploring smaller models. In addition, both the posterior model probability and constrained OLS will produce poorer BMA estimate, whereas in bigger model space the Gibbs sampler executes better. Therefore the corresponding BMA estimator calculated based on the models visited by Gibbs sampler, can be used as a tool to determine the nature of variables over the model space. In fact, in Section 6 we will use BMA estimator to compare the performance of FPE variable selection techniques.

3.1 Posterior Model Probability under Model (3)

Here we study in great detail the posterior model probability under model (3) as it is vital for model exploration. The posterior probability for α is

$$\Pr(\alpha|\mathbf{Y}) = \frac{\Pr(\mathbf{Y}|\alpha)\Pr(\alpha)}{\sum_{\alpha'} \Pr(\mathbf{Y}|\alpha')\Pr(\alpha')},$$

where $\Pr(\mathbf{Y}|\alpha) = \int f(\mathbf{Y}|\beta, \alpha) f(\beta|\alpha) d\beta$ is the marginal integrated likelihood of α and $\Pr(\alpha)$ is the marginal prior probability of model α. If we interpret $\Pr(\alpha|\mathbf{Y})$ as an assessment of the importance of model α, then from a model selection perspective; the goal is to choose α for which $\Pr(\alpha|\mathbf{Y})$ is largest. In the data analysis section (Section 6), we will present this probability for each variables in the model.

Theorem 1. Under the models (3) and (4), the posterior model probability of α is

$$\Pr(\alpha|\mathbf{Y}) = \frac{\Pr(\alpha)C(\alpha)}{\sum_{\alpha'} \Pr(\alpha')C(\alpha')},$$

where $C(\alpha) = |\hat{\sigma}^{-2}\mathbf{\Gamma}_\alpha\mathbf{\Sigma}_\alpha^{-1}|^{-1/2}\exp\{\boldsymbol{\mu}_\alpha^t\mathbf{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha/(2\hat{\sigma}^2)\}$, $\mathbf{\Sigma}_\alpha^{-1} = (\mathbf{X}^t\mathbf{X} + \hat{\sigma}^2\mathbf{\Gamma}_\alpha^{-1})$, $\mu_\alpha = \mathbf{\Sigma}_\alpha\mathbf{X}^t\mathbf{Y}$ is the conditional posterior mean of model α, $\mathbf{\Gamma}_\alpha$ is the $\alpha \times \alpha$ diagonal matrix $\text{diag}(\gamma_1, \cdots, \gamma_\alpha)$.

In describing the proposed bimodal model, V is set to a very large value, but the question is how large is this value? In Section 4, we come up with the exact value of V. In describing the Bayesian hierarchical model, we have mentioned that v_0 should be very small and V should be very large; so it is interesting to note the limiting form of the posterior model probability under the limits $v_0 \to 0$ and $V \to \infty$. Under our proposed model, the conditional posterior mean is approximately equivalent to constrained OLS estimator. The first Lemma will formally present this argument. The second Lemma will depict the limiting form of $C(\alpha)$. Combining both results, we present the limiting form of $\Pr(\alpha|\mathbf{Y})$ in the Theorem 2.

Lemma 1. As $v_0 \to 0$ and $V \to \infty$,

$$\boldsymbol{\mu}_\alpha \to \hat{\mu}_\alpha,$$

where $\hat{\mu}_\alpha$ denotes the estimator with values $\hat{\mu}_{+\alpha}$ along the α coordinates, and zero elsewhere, and $\hat{\mu}_{+\alpha} = (\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha})^{-1}\mathbf{X}_{+\alpha}^t\mathbf{Y}$ is the constrained OLS estimator for α.

Lemma 2. As $v_0 \to 0$ and $V \to \infty$,

$$C(\alpha) \xrightarrow{p} \hat{C}(\alpha),$$

where

$$\hat{C}(\alpha) = \left(\frac{\hat{\sigma}^2}{V}\right)^{K_{\alpha/2}} \mid \mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha} \mid^{-1/2} \exp\left\{\frac{1}{2\hat{\sigma}^2}\hat{\mu}_{+\alpha}^t[\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha}]\hat{\mu}_{+\alpha}\right\}.$$

Theorem 2. As $v_0 \to 0$ and $V \to \infty$, along with the results of Lemma 1 and Lemma 2, the posterior model probability of Theorem 1 is of the following form

$$\Pr(\alpha|\mathbf{Y}) \approx \frac{\Pr(\alpha)\hat{C}(\alpha)}{\sum_{\alpha'}\Pr(\alpha')\hat{C}(\alpha')},$$

where $\hat{\mu}_\alpha$ and $\hat{C}(\alpha)$ are as defined earlier. (Here " $\approx$ " stands for "approximately equivalent to").

The final form of the posterior model probability in Theorem 2; leads to a number of interesting issues with respect to model space exploration. The presence of a constrained OLS estimator; in the expression of the posterior model

probability is noticeable. We calculate the constrained OLS based on the indices of the model. With a smaller model, the constrained OLS will be calculated based on the indices of that model, which results in introducing a smaller posterior model probability. However, we do not want to restrict ourselves to only smaller models; but would like to visit a variety of models. This is not possible because of the presence of $V^{-K_\alpha/2}$; appearing in $\Pr(\alpha|\mathbf{Y})$ (posterior model probability) through $C(\alpha)$. In a diffuse prior; the probability will heavily penalize larger models, which results in a penalization of the posterior values. As a result, the Gibbs sampler based on (3) only visits smaller models.

4. Rescaled Spike and Slab Model: A Bimodal Spike and Slab Model

To resolve this setback; instead of using (3), we use a rescaled spike and slab model (Ishwaran and Rao, 2005b). This approach uses the same prior (4) for π, however, in place of Y_i in (3), we use the rescaled values $Y_i^* = \hat{\sigma}^{-1} n^{1/2} Y_i$. An appropriate adjustment to the variance is introduced to accommodate the rescaling. A rescaled spike and slab model defined in Ishwaran and Rao (2005b) is:

$$\begin{aligned} (Y_i^*|\mathbf{x}_i, \beta) &\overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{x}_i^t\beta, n), \quad i = 1, \cdots, n, \\ (\beta|\boldsymbol{\gamma}) &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}), \\ \boldsymbol{\gamma} &\sim \pi(d\boldsymbol{\gamma}). \end{aligned} \tag{8}$$

Note that this hierarchy includes a factor n; in the variance of the response to compensate for the rescaling. Also, any effect of σ^2 is removed from this model because the Y_i's are rescaled by $\hat{\sigma}^2$. For this reason, unlike in (3), the presence of $\hat{\sigma}^2$ is not needed.

By replacing $\mathbf{Y}$ by $\mathbf{Y}^*$ and setting $V = n$, the limiting posterior model probability under the limit $v_0 \to 0$ and $V \to \infty$ is

$$\hat{\Pr}(\alpha|\mathbf{Y}^*) \approx \frac{\Pr(\alpha)\hat{C}^*(\alpha)}{\sum_{\alpha'} \Pr(\alpha')\hat{C}^*(\alpha')}, \tag{9}$$

where

$$\hat{C}^*(\alpha) = |\mathbf{X}_{+\alpha}^t \mathbf{X}_{+\alpha}|^{-1/2} \exp\left\{ \frac{1}{2^2} \hat{\mu}_{+\alpha}^t [\mathbf{X}_{+\alpha}^t \mathbf{X}_{+\alpha}] \hat{\mu}_{+\alpha} \right\}. \tag{10}$$

(Here " $\approx$ " stands for "approximately equivalent to"). Perfect cancellation of V and $\hat{\sigma}^2$ occurring in $\hat{C}(\alpha)$, ensures that $\hat{C}^*(\alpha)$ is independent of V. Hence the issue of the unavoidable penalty term involved in the posterior model probability is resolved in (3).

We present some interesting results related to the posterior model probability under the bimodal model (8). The following theorem shows an interesting rela-

tionship between the posterior model probability (9) and a BIC-like penalization value.

Theorem 3. Under (8), maximizing $\hat{\Pr}(\alpha|\mathbf{Y}^*)$ is approximately equivalent to minimizing the BIC-penalty

$$\text{RSS}(\alpha) + \hat{\sigma}^2 K_\alpha \log(n). \tag{11}$$

The above theorem shows that there is an intimate association between the posterior for α and BIC-penalization, but to assume that model selection based on our method, is nothing more than a restricted all-subsets search guided by a BIC-like penalty, is incorrect and flawed. With the induction of the effects of the "true" complexity parameter (for details, see the proof of the Theorem 4 in the Appendix), we present further interesting results using the posterior mean in model (8). The following theorem demonstrates our viewpoint.

Theorem 4. Under (8), maximizing $\hat{\Pr}(\alpha|\mathbf{Y}^*)$ is approximately equivalent to minimizing

$$\text{RSS}(\alpha) + \hat{\sigma} K_\alpha \log(n) + 2\hat{\sigma}^2 K_\alpha \log\left(\frac{1-w_0}{w_0}\right). \tag{12}$$

This can be rewritten as

$$\text{RSS}(\alpha) + \text{BIC.pen}(\alpha) + \log\left(\frac{1-w_0}{w_0}\right) \times \text{AIC.pen}(\alpha), \tag{13}$$

where w_0 is the "true" complexity value, $\text{BIC.pen}(\alpha) = \hat{\sigma}^2 K_\alpha \log(n)$ and $\text{AIC.pen}(\alpha) = 2\hat{\sigma}^2 K_\alpha$ are the usual BIC and AIC penalty values.

The presence of w_0 in Theorem 4 leads to an interesting discussion on model exploration. The expression in (12) is not a traditional BIC penalty, instead a modified term involving the adaptively estimated complexity! For instance, a small value $0 < w_0 < 1/2$ will heavily penalize larger models, whereas a large $1/2 < w_0 < 1$ value will penalize smaller models. The value $w_0 = 1/2$ leads to a BIC-like penalty, since, setting $w_0 = 1/2$ in (12) yields (11).

In fact, one can view (12) or (13) as a compromise between BIC and an adaptively estimate of the AIC penalty. This is another way to interpret the effect of w_0. Setting $w_0 = 1/2$ in (12) or (13) eliminates the AIC penalty. On the other hand, as w_0 approaches 1, the AIC penalization term becomes negative and the overall penalization drops, which encourages larger models. The opposite occurs as w_0 becomes small, as w_0 approaches 0; it results in a positive and very large penalization; yielding small models.

5. Simulation Study

5.1 Example 1

A simulation study is performed to compare the performance of this model exploration tool compared to other popular model selection methods. we are using "Breiman simulations" as described in Breiman (1992). In the simulations, data is generated with the following setting: the covariates $\mathbf{x}_i$ are simulated independently from a multivariate normal distribution with $E(x_{i,k}) = 0$ and $E(x_{i,j}x_{i,k}) = \rho^{|j-k|}$, where $0 < \rho < 1$ represent the correlation coefficient and the error ϵ_i's are i.i.d $N(0,1)$ variables. Three different conditions of ρ are considered: (i) uncorrelated design ($\rho = 0$), (ii) moderate correlated design ($\rho = 0.5$), and (iii) highly correlated design ($\rho = 0.9$).

Under this global condition, two different conditions are considered: **(I)** a model with moderate number of non-zero coefficients ($n = 200$, $K = 100$, there are 55 zero coefficients); and **(II)** a model with many zero coefficients ($n = 800$, $K = 400$, there are 295 zero coefficients). For model **(I)**, non-zero coefficients are clustered into 9 different groups, each cluster consists of 5 coefficients. Value of these 5 coefficients remain unchanged and replicates across the entire clusters. For model **(II)**, non-zero coefficients are clustered into 15 different groups, each cluster consists of 7 coefficients. Value of these 7 coefficients remain unchanged and replicates across the entire clusters. In all simulated models, coefficient values are adjusted by multiplying a constant by which the theoretical R^2 is 0.75.

The goal of this simulation study is to see how well our model exploration tool performs in comparison to lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006) and relaxed lasso (Meinshausen, 2007). Relaxed lasso is an alternative to lasso; where a double optimization technique is implemented using two tuning parameters to improve variable selection performance, for details see Meinshausen (2007). The adaptive lasso is another modification to lasso, works different than relaxed lasso, adds a unique penalization to each coefficient on top of lasso penalization, see Zou (2006) for details. Most of the time these methods are used for model selection; here we use these methods from the perspective of model space exploration. For this reason, we focus on a couple of items: estimated model size ($\hat{k}$) and total number of incorrectly classified variables (Miss). Lasso computation is implemented by using R package lars; relaxed lasso is implemented by using R package relaxo. Tables 1 and 2 summarizes the simulation results.

Below is the summary of our findings from the simulation study (with reference to Tables 1 and 2).

Bimodal model always opts for smaller model. The lasso selects larger models and has a tendency to overfit the model. In certain situations; misclassification rate for bimodal model is lesser than in lasso and other methods. In most situations; relaxed lasso performs quite identical to lasso. In correlated setting, performance of adaptive lasso is inferior than lasso as far as misclassification rate

Table 1: Simulation result based on model **(I)** of Example 1. The values used in the simulation are: sample size (n) = 200, total number of variables (K) in the model = 100, out of them 55 (55%) are zero coefficients. Reported criteria are $\hat{k}$ (averaged estimated model size), Correct (average number of correctly classified variables), and Miss (average number of misclassified variables). Monte Carlo standard deviations are in parentheses below each criterion. All results are based on 200 iterations

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
	$\hat{k}$	Correct	Miss	$\hat{k}$	Correct	Miss	$\hat{k}$	Correct	Miss
bimodal	29.93	25.67	22.23	19.67	17.88	25.54	11.42	9.41	37.63
	(3.79)	(2.89)	(2.61)	(3.18)	(1.92)	(1.92)	(1.10)	(0.98)	(1.48)
lasso	67.1	41.08	30.68	50.36	36.26	22.94	32.43	21.34	34.31
	(7.15)	(1.91)	(5.98)	(7.89)	(2.92)	(4.81)	(5.91)	(3.12)	(4.67)
adaptive lasso	37.11	32.64	18.41	28.04	24.04	26.44	17.18	11.20	39.74
	(6.05)	(3.88)	(3.55)	(6.19)	(3.59)	(3.86)	(2.88)	(3.15)	(3.28)
relaxed lasso	71.3	39.15	34.04	49.71	32.54	23.99	34.11	19.68	35.33
	(14.33)	(3.56)	(9.70)	(13.25)	(4.56)	(7.70)	(9.92)	(3.58)	(5.85)

Table 2: Simulation result based on model **(II)** of Example 1. The values used in the simulation are: sample size (n) = 800, total number of variables (K) in the model = 400, out of them 295 (74%) are zero coefficients. Reported criteria are $\hat{k}$ (averaged estimated model size) , Correct (average number of correctly classified variables), and Miss (average number of misclassified variables). Monte Carlo standard deviations are in parentheses below each criterion. All results are based on 200 iterations

	$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
	$\hat{k}$	Correct	Miss	$\hat{k}$	Correct	Miss	$\hat{k}$	Correct	Miss
bimodal	63.79	57.67	49.98	42.28	39.14	66.57	22.10	19.45	81.44
	(3.78)	(3.27)	(3.08)	(3.10)	(2.39)	(2.15)	(1.43)	(1.03)	(1.15)
lasso	196.61	92.92	114.77	133.83	84.86	69.71	83.78	58.77	70.14
	(16.89)	(4.27)	(14.54)	(16.65)	(3.59)	(13.18)	(8.67)	(3.97)	(7.67)
adaptive lasso	86.31	73.98	42.91	68.03	55.38	60.57	47.88	29.14	92.76
	(8.65)	(4.86)	(4.72)	(8.23)	(5.69)	(4.41)	(6.97)	(3.32)	(5.27)
relaxed lasso	162.9	84.10	91.84	101.26	72.46	51.76	83.59	52.75	71.25
	(55.69)	(7.21)	(41.92)	(30.25)	(7.96)	(18.90)	(16.89)	(6.59)	(12.44)

is concerned. An interesting feature of the bimodal model all through the simulation study is that the Monte Carlo standard deviations are minimal in comparison to other methods, this explains the stability of the bimodal model in exploring the model space.

In all different conditions, for bimodal model, based on the estimated model size, the number of correctly specified non-zero variables on an average; perform reasonably well with respect to other comparable methods.

Finally, it is seen that in certain circumstances; misclassification rate is higher for bimodal model than other methods as the model size is smaller compared to other methods. Here lies the trade-off between misclassification rate and model size: if the model size is bigger, misclassification rate is low and vice versa.

5.2 Example 2

Here we conduct another simulation study to investigate the performance of the proposed bimodal model using high-dimensional sparse linear model as used in Ing and Lai (2011) and Shao and Chow (2007). Two different sets of sample size and number parameters are considered in the model: $(n = 50, K = 200)$ and $(n = 100, K = 400)$. For each of the above condition, error standard deviations are taken as $\sigma = 1$ and $\sigma = 0.1$. It is assumed that only the first five covariates are non-zero and rest of them are zero.

The simulated data is generated as follows. The response Y_i's are generated as $Y_i = \mathbf{x}_i^t\beta + \epsilon_i,\ i = 1, 2, \cdots, n$. The set of $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ are independently generated from a multivariate normal distribution with mean vector $\mathbf{1}_K$ (the K-dimensional vector of ones) and covariance matrix $\mathbf{I}_K$ (an identity matrix of order K). Two different constraints on $\mathbf{x}_i$'s are considered: $\mathbf{x}_i$'s are fixed throughout the iterations and $\mathbf{x}_i$'s are random throughout the iterations. The β's values are taken as $\beta = (3, -3.5, 4, -2.8, 3.2, 0, 0, \cdots, 0)^t$. The error ϵ_i's are independently generated from $N(0, \sigma^2)$.

Table 3 reports the performance of the bimodal spike and slab model in high-dimensional sparse linear model. In all different situations, bimodal model shows quite similar trend in variable selection, correctly specifying non-zero variables and also in incorrectly specifying variables. There is not much difference noticed in variations of model properties. In all the different situations, bimodal model produces smaller model size. There are only five non-zero covariates in the model; from this study it is noticeable that the correctly identified non-zero covariates are approximately two on an average, which is fairly fitting, and the misclassification rate is also reasonably effective. It should be noted that in current model/variable selection techniques for high-dimensional sparse models, most of the time a two-stage procedure has been used; first a variable screening technique is used to reduce the number of variables from the model; then a second method has been used to the reduced model for final model/variable selection. Here we don't use any screening method; we directly use the bimodal model for variable selection.

Table 3: Simulation result based on Example 2. Reported criteria are $\hat{k}$ (averaged estimated model size), Correct (average number of correctly classified variables), and Miss (average number of misclassified variables). All results are based on 200 iterations

σ	n	K	$\hat{k}$	Correct	Miss
Fixed **x**					
1	50	200	12.73	2.30	13.13
	100	400	26.09	1.84	28.39
0.1	50	200	19.96	1.86	21.63
	100	400	23.76	1.73	25.31
Random **x**					
1	50	200	12.87	1.85	14.17
	100	400	19.12	1.79	22.73
0.1	50	200	13.96	1.65	15.66
	100	400	21.36	1.69	20.78

6. Applications of Bimodal Spike and Slab Model

6.1 Ozone Data

This section applies our R software package modelSampler for variable selection by FPE method using the well known ozone data (Breiman and Friedman, 1985). The data set consists of 203 observations with 12 variables, each observation is a day. The outcome variable in this example is ozone emission, ozone.

Table 4 contains the variable selection result based on ozone data. The first column (variable names) is ordered by the BMA estimators. The second column represents the posterior inclusion probability values of those variables. If the goal is to perform variable selection based on the median model (Barbieri and Berger, 2004) (0.50 of "prob" value is the cut-off point for selecting a variable to be in the model), then according to the median model, only four variables are in the selected model. The last two columns indicate the AIC and BIC models. The table shows that six variables are selected by AIC whereas only three variables are selected by BIC. It is interesting to note that the BIC model is more conservative than the median model.

Table 5 helps to assess the importance of a variable. Variables in the first column of the table are ordered by the BMA estimators, as a result the ordering is quite stable. This indicates the best model of size k for each k in terms of R^2 value. This table allows one to weigh in the global importance of a given variable. For example, the variable "humid" is highly significant as it appears in the best model for any given size. Also the posterior inclusion probability for "humid" is

Table 4: Variable selection results based on Ozone data. Results are based on a 2,500 burn-in iteration and 10,000 sampled values

Variable	prob	median model	AIC model	BIC model
tempE	0.819	✓	✓	✓
humid	0.981	✓	✓	✓
tempS	0.610	✓	✓	
month	0.903	✓	✓	✓
ibtemp	0.276			
press	0.287		✓	
ibht	0.263		✓	
vis	0.140			
pressg	0.114			
daymonth	0.062			
wind	0.063			
dayweek	0.056			

Table 5: Top model stratified by size of the model based on Ozone data. Results are based on a 2,500 burn-in iteration and 10,000 sampled values

Variable	1	2	3	4	5	6	7	8	9	10	11	12
tempE		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
humid	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
tempS					✓	✓	✓	✓	✓	✓	✓	✓
month			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ibtemp								✓	✓	✓	✓	✓
press				✓	✓	✓	✓	✓	✓	✓	✓	✓
ibht						✓	✓	✓	✓	✓	✓	✓
vis							✓	✓	✓	✓	✓	✓
pressg											✓	✓
daymonth											✓	✓
wind									✓	✓	✓	✓
dayweek										✓		✓

0.98, the largest value of all the variables. This result demonstrates the behavior of the variables over the model space.

6.2 Graphical Diagnostics

Here we graphically summarize the results obtained from the analysis using a graphical wrapper which is produced from the R package (see Figure 1). The wrapper consists of five distinct plots. The top leftmost plot represents a histogram of several estimated complexity parameters from 10,000 sampled models. A closer look at the plot indicates that the mode of the complexity value lies in

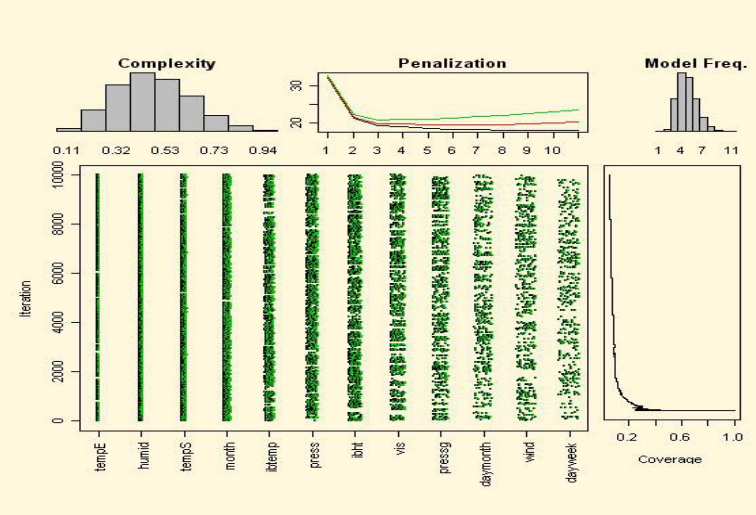

Figure 1: Complete graphical analysis of Ozone data. Results are based on a 2,500 burn-in iteration and 10,000 sampled values

between 0.33 to 0.56. The middle plot on the top represents FPE values as a function of the dimension of the models. The black line corresponds to minimum residual sum of squares, the red line is minimum AIC and the green line is minimum BIC. The rightmost plot on the top is a histogram of several dimensionality of the models visited by the Gibbs sampler. A model of size 3 to 5 is the most frequently visited model. This characterizes the fact that three to five variables are the most significant variables in the data set. The bottom leftmost plot is an interesting plot. Here we have plotted the variables of the sampled model with respect to the number of Monte Carlo iterations. This plot helps to re-evaluate the importance of variables implicitly. For example, humid is present in almost all sampled models, while dayweek hardly ever appears. Clearly humid is an important variable. The last plot on the bottom represents the probability of visiting a new model at each iteration. We see that it stabilizes very quickly as the line is almost flat as the number of iterations approaches 10,000; which graphically identifies the convergence of the Gibbs sampler.

These plots are highly effective in explaining the data. Model space exploration helps to understand the nature of variables in explaining the data. A closer look at the plots show strong evidence of our argument that without even implementing any model selection technique; we have a clear-cut idea of the relevant variables and the data dimensionality. The leftmost plot on the bottom is a perfect example in this situation. As we sample several models, it is clear from this plot, which all variables are all-encompassing of any model.

7. Discussion

In this article we propose a simple bimodal model under spike and slab hierarchy. From the model space exploration point of view; this model is appropriate as it has selective shrinkage property. We have discussed the interplay between conditional posterior mean and model space exploration. Theoretical justifications illustrate the asymptotic equivalence between conditional posterior mean and constrained OLS estimator, that results in implementing frequentist model selection approach. Another interesting feature of bimodal prior is that it associates with the popular g-prior introduced by Zellner (1986), see Dey (2008) for details. By conducting an extensive simulation study; we show that the bimodal model favorably prevails against lasso, adaptive lasso and relaxed lasso. In all simulations it generated smaller models with competing misclassification rate. The performance of the proposed method has also been tested in high-dimensional sparse linear model via simulation study. The bimodal model shows quite reasonable performance by producing smaller models and with acceptable misclassification rate and correctly classifying non-zero covariates in the model. The proposed method is effective in both model space exploration and variable selection process by using both frequentist and Bayesian techniques. The R package modelSampler is available upon request.

Appendix

Proof of the Theorem 1. First note that

$$f(\mathbf{Y}|\beta,\alpha)f(\beta|\alpha) = (2\pi\hat{\sigma}^2)^{-n/2}|2\pi\mathbf{\Gamma}_\alpha|^{-1/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}||\mathbf{Y}||^2 + \frac{1}{2\hat{\sigma}^2}\boldsymbol{\mu}_\alpha^t\mathbf{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha\right\}$$
$$\times\exp\left\{-\frac{1}{2\hat{\sigma}^2}(\beta-\boldsymbol{\mu}_\alpha)^t\mathbf{\Sigma}_\alpha^{-1}(\beta-\boldsymbol{\mu}_\alpha)\right\}.$$

Therefore,

$$\begin{aligned}\Pr(\mathbf{Y}|\alpha) &= \int f(\mathbf{Y}|\beta,\alpha)f(\beta|\alpha)d\beta\\ &= (2\pi\hat{\sigma}^2)^{-n/2}|2\pi\mathbf{\Gamma}_\alpha|^{-1/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}||\mathbf{Y}||^2 + \frac{1}{2\hat{\sigma}^2}\boldsymbol{\mu}_\alpha^t\mathbf{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha\right\}\\ &\quad\times\int\exp\left\{-\frac{1}{2\hat{\sigma}^2}(\beta-\boldsymbol{\mu}_\alpha)^t\mathbf{\Sigma}_\alpha^{-1}(\beta-\boldsymbol{\mu}_\alpha)\right\}d\beta\\ &= (2\pi)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}||\mathbf{Y}||^2\right\}C(\alpha),\end{aligned}$$

where

$$C(\alpha) = |\hat{\sigma}^{-2}\mathbf{\Gamma}_\alpha\mathbf{\Sigma}_\alpha^{-1}|^{-1/2}\exp\left\{\frac{1}{2\hat{\sigma}^2}\boldsymbol{\mu}_\alpha^t\mathbf{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha\right\}.$$

Consequently,

$$\Pr(\alpha|\mathbf{Y}) = \frac{\Pr(\mathbf{Y}|\alpha)\Pr(\alpha)}{\sum_{\alpha'}\Pr(\mathbf{Y}|\alpha')\Pr(\alpha')} = \frac{C(\alpha)\Pr(\alpha)}{\sum_{\alpha'}C(\alpha')\Pr(\alpha')}$$

is the posterior model probability for α. □

Proof of Lemma 1.

$$\lim_{V\to\infty}\boldsymbol{\mu}_\alpha = \lim_{V\to\infty}\left\{(\mathbf{X}^t\mathbf{X} + \hat{\sigma}^2\mathbf{\Gamma}_\alpha^{-1})^{-1}\mathbf{X}^t\mathbf{Y}\right\} = \left(\mathbf{X}^t\mathbf{X} + \frac{\hat{\sigma}^2}{v_0}(\mathbf{I} - \mathbf{I}_\alpha^0)\right)^{-1}\mathbf{X}^t\mathbf{Y},$$

where $\mathbf{I}_\alpha^0$ is the diagonal matrix with a zero entry if $k \notin \alpha$ and an entry of 1 if $k \in \alpha$. Hereafter, we use a subscript of $+\alpha$ to indicate the inclusion of only those values in α and a subscript of $-\alpha$ to indicate the exclusion of all values in α. For the moment, without loss of generality, assuming that $\mathbf{X}$ is ordered so that the first K_α coordinates correspond to α. Using the following partition of A:

$$\mathbf{A} = \left(\mathbf{X}^t\mathbf{X} + \frac{\hat{\sigma}^2}{v_0}(\mathbf{I} - \mathbf{I}_\alpha^0)\right) = \begin{pmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{1,2}^t & v_0^{-1}\mathbf{A}_{2,2} \end{pmatrix},$$

where

$$\mathbf{A}_{1,1} = \mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha},\ \ \mathbf{A}_{1,2} = \mathbf{X}_{+\alpha}^t\mathbf{X}_{-\alpha},\ \text{ and }\ \mathbf{A}_{2,2} = v_0\mathbf{X}_{-\alpha}^t\mathbf{X}_{-\alpha} + \hat{\sigma}^2\mathbf{I}_{-\alpha},$$

validating that

$$\mathbf{A}^{-1} = \mathbf{B} = \begin{pmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} \\ \mathbf{B}_{1,2}^t & \mathbf{B}_{2,2} \end{pmatrix},$$

where

$$\begin{aligned}
\mathbf{B}_{1,1} &= \left(\mathbf{A}_{1,1} - v_0\mathbf{A}_{1,2}\mathbf{A}_{2,2}^{-1}\mathbf{A}_{2,1}\right)^{-1}, \\
\mathbf{B}_{1,2} &= -v_0\mathbf{B}_{1,1}\mathbf{A}_{1,2}\left(\mathbf{A}_{2,2} - v_0\mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\right), \\
\mathbf{B}_{2,2} &= v_0\left(\mathbf{A}_{2,2} - v_0\mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\right)^{-1}.
\end{aligned}$$

If $v_0 \to 0$, then $\mathbf{A}_{2,2} \to \hat{\sigma}^2\mathbf{I}_{-\alpha}$, an invertible matrix. Deducing

$$\mathbf{A}^{-1} \to \begin{pmatrix} \mathbf{A}_{1,1}^{-1} & \mathbf{0}_{+\alpha,-\alpha}, \\ \mathbf{0}_{+\alpha,-\alpha}^t & \mathbf{0}_{-\alpha,-\alpha}, \end{pmatrix},\ \text{ as }\ v_0 \to 0.$$

Therefore,

$$\boldsymbol{\mu}_\alpha \to \hat{\mu}_\alpha, \quad \text{as} \quad v_0 \to 0 \quad \text{and} \quad V \to \infty.$$

□

Proof of Lemma 2. Let $\boldsymbol{\Gamma}_\alpha^0 = \left[\mathbf{I}_\alpha^0 + (v_0/V)(\mathbf{I} - \mathbf{I}_\alpha^0)\right]^{1/2}$. Observing that

$$\begin{aligned} V^{K/2}|\hat{\sigma}^{-2}\boldsymbol{\Gamma}_\alpha\boldsymbol{\Sigma}_\alpha^{-1}|^{-1/2} &= \hat{\sigma}^K|\boldsymbol{\Gamma}_\alpha^0\mathbf{X}^t\mathbf{X}\boldsymbol{\Gamma}_\alpha^0 + \hat{\sigma}^2V^{-1}\mathbf{I}|^{-1/2} \\ &\to \hat{\sigma}^K|\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha} + \hat{\sigma}^2V^{-1}\mathbf{I}_{+\alpha}|^{-1/2} \times |\mathbf{I}_{-\alpha}|^{-1/2} \times \left(\frac{\hat{\sigma}^2}{V}\right)^{-(K-K_\alpha)/2}, \\ &\text{as} \quad v_0 \to 0. \end{aligned}$$

Assuming that if $V \to \infty$,

$$C(\alpha) \xrightarrow{p} \hat{C}(\alpha), \quad \text{as} \quad v_0 \to 0.$$

□

Proof of Theorem 3. By multiplying the numerator and denominator of $\hat{\Pr}(\alpha|\mathbf{Y}^*)$ by $\exp(-||\mathbf{Y}||^2/(2\hat{\sigma}^2))$, we replace (10) with

$$\hat{C}^*(\alpha) = |\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha}|^{-1/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\mathrm{RSS}(\alpha)\right\}, \tag{14}$$

where $\mathrm{RSS}(\alpha) = ||\mathbf{Y} - \mathbf{X}_{+\alpha}\hat{\mu}_{+\alpha}||^2$ is the residual sum-of-squares for α. This follows from the identity

$$-\mathrm{RSS}(\alpha) = -||\mathbf{Y}||^2 + \mathbf{Y}^t\mathbf{X}_{+\alpha}\hat{\mu}_{+\alpha} = -||\mathbf{Y}||^2 + \hat{\mu}_{+\alpha}^t[\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha}]\hat{\mu}_{+\alpha}.$$

The first term on the right of (14) is a dimensionality effect. On the log-scale this is roughly of order $-K_\alpha\log(n)/2$. For example, if $\mathbf{X}$ is orthogonal, then assuming rescaling of covariates, $\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha} = n\mathbf{I}_{+\alpha}$. This rescaling is crucial, since it implies that

$$|\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha}|^{-1/2} = n^{-K_\alpha/2}. \tag{15}$$

The log of the second term on the right of (14) is a measure of goodness of fit, $-\mathrm{RSS}(\alpha)/(2\hat{\sigma}^2)$. Thus, maximizing $\hat{C}^*(\alpha)$ is equivalent to minimizing the BIC-penalty

$$\mathrm{RSS}(\alpha) + \hat{\sigma}^2K_\alpha\log(n).$$

□

Proof of Theorem 4. First observe that (9) can be rewritten as

$$\hat{\Pr}(\alpha|\mathbf{Y}^*) = \int_0^1 \frac{\Pr(\alpha|w)\hat{C}^*(\alpha)}{\sum_{\alpha'}\Pr(\alpha'|w)\hat{C}^*(\alpha')}\,\nu(dw|\mathbf{Y}^*),$$

where

$$\nu(dw|\mathbf{Y}^*) = \frac{\nu(dw)\sum_{\alpha'}\Pr(\alpha'|w)\hat{C}^*(\alpha')}{\sum_{\alpha'}\Pr(\alpha')\hat{C}^*(\alpha')}.$$

(This follows either by direct application of Bayes theorem, or upon substituting the expression for $\nu(dw|\mathbf{Y}^*)$ and integrating). In particular ν is a prior measure for $w \in [0,1]$. Now

$$\Pr(\alpha|w) = w^{K_\alpha}(1-w)^{K-K_\alpha}$$

is the prior probability for α given w. Because we expect $\nu(dw|\mathbf{Y}^*) \to w_0$, the "true" complexity value, this implies

$$\hat{\Pr}(\alpha|\mathbf{Y}^*) \approx \frac{\Pr(\alpha|w_0)\hat{C}^*(\alpha)}{\sum_{\alpha'}\Pr(\alpha'|w_0)\hat{C}^*(\alpha')},$$

and consequently:

$$\begin{aligned}\log\Big(\hat{\Pr}(\alpha|\mathbf{Y}^*)\Big) \;\approx\; & K_\alpha\log(w_0) + (K-K_\alpha)\log(1-w_0) \\ & -\frac{1}{2}\log\Big(|\mathbf{X}_{+\alpha}^t\mathbf{X}_{+\alpha}|\Big) - \frac{1}{2\hat{\sigma}^2}\mathrm{RSS}(\alpha) + \text{Constant}.\end{aligned}$$

(Here " $\approx$ " stands for "approximately equivalent to"). Using a similar argument as in the proof of Theorem 3, maximizing this value is equivalent (approximately) to minimizing

$$\mathrm{RSS}(\alpha) + \hat{\sigma}^2 K_\alpha \log(n) + 2\hat{\sigma}^2 K_\alpha \log\left(\frac{1-w_0}{w_0}\right).$$

The above expression can be rewritten as

$$\mathrm{RSS}(\alpha) + \mathrm{BIC.pen}(\alpha) + \log\left(\frac{1-w_0}{w_0}\right) \times \mathrm{AIC.pen}(\alpha).$$

□

Acknowledgements

The author would like to thank the referee and the editors who helped to substantially improve the paper. The author would also like to thank Ross Iaci and David Lutzer for prudent reading and helpful suggestions to enhance the quality of the paper.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243-247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Czàki), 267-281. Akademiai Kiadò, Budapest.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 870-897.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**, 580-598.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738-754.

Dey, T. (2008). *Prediction and Variable Selection.* Ph.D. Dissertation, Department of Statistics, Case Western Reserve University.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499-511.

George, E. L. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

George, E. L. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382-417.

Ing, C. K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21**, 1473-1513.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438-455.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100**, 764-780.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730-773.

Ishwaran, H. and Rao, J. S. (2011). *Generalized Ridge Regression*: *Geometry and Computational Solutions when p is Larger than n.* Technical report, Division of Biostatistics, University of Miami.

Lumley, T. (2009). *Leaps*: *Regression Subset Selection.* R package version 2.7. http://cran.r-project.org/web/packages/leaps

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis* **52**, 374-393.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023-1036.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I. and Yeung, K. Y. (2009). *BMA*: *Bayesian Model Averaging.* R package version 3.10. http://CRAN.R-project.org/package=BMA

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning* **5**, 197-227.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

Shao, J. and Chow, S. C. (2007). Variable screening in predicting clinical outcome with high-dimensional microarrays. *Journal of Multivariate Analysis* **98**, 1529-1538.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques - Essays in Honor of Bruno de Finetti* (Edited by P. K. Goel and A. Zellner), 233-243. Amsterdam, North-Holland.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Received June 13, 2011; accepted December 13, 2011.

Tanujit Dey
Department of Mathematics
College of William & Mary
Williamsburg, VA 23185, USA
tdey@wm.edu