The Annals of Statistics
2011, Vol. 39, No. 3, 1748–1775
DOI: 10.1214/11-AOS884

A MAJORIZATION–MINIMIZATION APPROACH TO VARIABLE SELECTION USING SPIKE AND SLAB PRIORS

BY TSO-JUNG YEN[1]

Academia Sinica

We develop a method to carry out MAP estimation for a class of Bayesian regression models in which coefficients are assigned with Gaussian-based spike and slab priors. The objective function in the corresponding optimization problem has a Lagrangian form in that regression coefficients are regularized by a mixture of squared l_2 and l_0 norms. A tight approximation to the l_0 norm using majorization–minimization techniques is derived, and a coordinate descent algorithm in conjunction with a soft-thresholding scheme is used in searching for the optimizer of the approximate objective. Simulation studies show that the proposed method can lead to more accurate variable selection than other benchmark methods. Theoretical results show that under regular conditions, sign consistency can be established, even when the Irrepresentable Condition is violated. Results on posterior model consistency and estimation consistency, and an extension to parameter estimation in the generalized linear models are provided.

1. Introduction.

Consider the following regression model:

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i, \tag{1.1}$$

where Y_i is the response variable for the ith subject, x_{ij} is the jth covariate for the ith subject, β_j is the corresponding regression coefficient and ε_i is the error term following some specified distribution. Variable selection in regression problems has long been considered as one of the most important issues in modern statistics. It involves choosing an appropriate subset $\widehat{S}$ of indices $\{1, 2, \ldots, p\}$ so that for $j \in \widehat{S}$, the covariates x_{ij}'s and estimated coefficients $\widehat{\beta}_j$'s are scientifically meaningful in interpretation, and estimates $\widehat{y}_{i'} = \sum_{j\in\widehat{S}} x_{i'j}\widehat{\beta}_j$ have relative good properties in prediction.

In this paper, we develop a method to carrying out maximum a posteriori (MAP) estimation for a class of Bayesian models in tackling variable selection problems. The use of MAP estimation in variable selection problems had previously been studied by Genkin et al. [11] on logistic regression models with Laplace priors.

Received July 2010; revised December 2010.

[1] Supported in part by Grants NSC 97-3112-B-001-020 and NSC 98-3112-B-001-027 in the National Research Program for Genomic Medicine.

MSC2010 subject classifications. Primary 62H12; secondary 62F15, 62J05.

Key words and phrases. MAP estimation, l_0 norm, majorization–minimization algorithms, Irrepresentable Condition.

The difference between our model and Genkin et al.'s is that our model assigns a Gaussian-based spike and slab prior weighted by Bernoulli variables on each regression coefficient. Traditionally, parameter estimation for this model and other Bayesian variable selection settings relies on Markov chain Monte Carlo for posterior simulation [13, 14, 16, 17, 23, 32] and empirical Bayes methods [4, 12, 21]. A major advantage of MCMC-based inference procedures is that they provide a practical way to assessing posterior probabilities, and inference tasks such as point estimation can be carried out straightforwardly based on posterior probability calculation. However, convergence of MCMC-based sampling algorithms is not often guaranteed and they may become time-consuming as the number of covariates p becomes quite large. A different inference procedure on models with spike and slab priors is recently provided by Ishwaran and Rao [19, 20], in that regression coefficients are estimated via OLS-based shrinkage methods.

Our estimation method is different from the above approaches in several aspects. In our MAP estimation, an augmented version of the posterior joint density is derived. From frequentists' point of view, the MAP estimation is equivalent to the regularization estimation with a mixture penalty of squared l_2 and l_0 norms on regression coefficients. In practice, we apply a majorization–minimization technique to modify the penalty function so that convexity of the objective function can be achieved. We then construct a coordinate descent algorithm based on a specified iteration scheme to obtain the MAP estimate. The algorithm involves iteratively applying shrinkage-thresholding steps to obtain estimates that have sparse features, that is, some of them have exact zero values. In this sense, parameter estimation and variable selection can be achieved simultaneously. In addition, the algorithm can be implemented practically in a situation in which the number of covariates p is much larger than the number of samples n. It is different from the OLS-based methods in that $p \leq n$ is required to avoid singularity in matrix operation. The algorithm can also be fast when p is large but the number of covariates with nonzero coefficients is small. Simulation studies show that the MAP estimate can lead to better performances in variable selection than those based on other benchmark methods in various circumstances.

Recent frequentists' approaches to variable selection focus on applying the idea of regularization estimation in the situation in which the number of variables is much larger than the number of samples [2, 7, 10, 27, 28, 36, 38, 40–43]. All these approaches can either been seen as alternatives or as extensions of the lasso estimation [33]. For theoretical properties of the lasso estimation, Knight and Fu [22] pointed out that with regular conditions on the order magnitude of the tuning parameter, the lasso is consistent in parameter estimation. However, as shown by Meinshausen and Bühlmann [29] and Zou [40], for the lasso estimation, consistency in parameter estimation does not imply consistency in variable selection. Further conditions on the design matrix and tuning parameter should be imposed to ensure consistency in variable selection for the lasso estimation. In this aspect, Zhao and Yu [39] established the Irrepresentable Condition and showed that the lasso can be asymptotically consistent in both variable selection and parameter estimation if the Irrepresentable Condition holds and some regular conditions on the tuning parameter are satisfied. The same condition was also established by Zou [40] and Yuan and Lin [37]. Later we will show that the MAP estimator proposed in this paper is asymptotically consistent in variable selection even when frequentists' Irrepresentable Condition is violated.

The paper is organized as follows. Section 3 focuses methodological aspects of the proposed method. Section 4 provides two simulation studies on performances of the proposed method. Section 5 develops relevant asymptotic analysis for the method. Section 6 extends the method to parameter estimation in the generalized linear models. Real data examples are provided in Section 7. Some concluding remarks are given in Section 8.

2. Notation.

Let X be an $n \times p$ design matrix. Let x_i denote the ith row of X and x_{ij} denote the ijth entry of X. The transpose of X is denoted by X^T. Let $y = (y_1, y_2, \dots, y_n)$ denote the realization of random vector $Y = (Y_1, Y_2, \dots, Y_n)$ and $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ denote the regression coefficient vector. Let $I_{p\times p}$ denote the $p \times p$ identity matrix. For a p-dimensional vector $a = (a_1, a_2, \dots, a_p)$, define the l_1 norm by $\|a\|_1 = \sum_{j=1}^p |a_j|$, the l_2 norm by $\|a\|_2 = (\sum_{j=1}^p |a_j|^2)^{1/2}$, the l_∞ norm by $\|a\|_\infty = \max_j |a_j|$ and the l_0 norm by $\|a_j\|_0 = \sum_{j=1}^p \mathbb{I}(a_j \neq 0)$, where $\mathbb{I}(a_j \neq 0)$ is an index variable such that $\mathbb{I}(a_j \neq 0) = 1$ if $a_j \neq 0$ and $\mathbb{I}(a_j \neq 0) = 0$ otherwise. The probability density of a random variable Z conditional on θ is denoted by $f(z|\theta)$. We define $S = \{j : \beta_j \neq 0, j = 1, 2, \dots, p\}$, that is, the index set of nonzero valued coefficients in $\beta = (\beta_1, \beta_2, \dots, \beta_p)$. We further define X_S as the design matrix of X whose columns are indexed by S. Finally, we define the sign function for variable z as $\text{sign}(z) = 1$ if $z > 0$; $\text{sign}(z) = -1$ if $z < 0$; $\text{sign}(z) = 0$ if $z = 0$.

3. The method.

We start by assigning prior distributions on parameters in the regression model (1.1). Note that in a regression model a covariate can only be selected if its coefficient is estimated with a nonzero value. Based on this observation, we assign an index variable γ_j to each covariate and define that $\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ if $\beta_j = 0$. Here, we may write $\gamma_j = \mathbb{I}(\beta_j \neq 0)$. With the definition of γ_j, the regression model (1.1) has an equivalent representation given by

$$Y_i = \sum_{j=1}^{p} x_{ij}\gamma_j\beta_j + \varepsilon_i.$$

From a variable selection point of view, the index vector $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)$ is an indicator for candidate models. Different candidate models will have different values in γ.

3.1. *The Bayesian formulation.* Under a Bayesian framework, we assume

$$
\begin{aligned}
&Y_i|x_i, \beta, \gamma, \sigma^2 \sim \text{Normal}\left(\sum_{j=1}^{p} x_{ij}\gamma_j\beta_j, \sigma^2\right) \qquad \text{for } i = 1, 2, \ldots, n,\\
&\beta_j|\sigma^2, \gamma_j, \lambda \sim \gamma_j \,\text{Normal}(0, \sigma^2\lambda^{-1}) + (1-\gamma_j)\mathbb{I}(\beta_j = 0)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } j = 1, 2, \ldots, p,\\
&\sigma^2|\tau_1, \tau_2 \sim \text{Inverse-Gamma}(\tau_1, \tau_2),\\
&\gamma_j|\kappa \sim \text{Bernoulli}(\kappa) \qquad \text{for } j = 1, 2, \ldots, p.
\end{aligned}
\tag{3.1}
$$

The prior distribution of β_j given in (3.1) is the spike and slab prior originally proposed by Mitchell and Beauchamp [30]. It implies that conditional on $\gamma_j = 0$, β_j is equal to 0 with probability one, and conditional on $\gamma_j = 1$, β_j follows a normal distribution with mean 0 and variance $\sigma^2\lambda^{-1}$. The Bernoulli prior on γ_j says that if only prior information is available, γ_j will have probability κ to be 1 and $1-\kappa$ to be 0. Note that since $\gamma_j \in \{0, 1\}$, we can express the mixture form of the prior on β_j as $\text{Normal}(0, \sigma^2\lambda^{-1})^{\gamma_j} \times \mathbb{I}(\beta_j = 0)^{1-\gamma_j}$. This representation will be used in deriving the joint posterior density of the parameters.

Under Bayesian model (3.1), the joint posterior density of β, γ and σ^2 can be expressed as

$$
\begin{aligned}
&f(\beta, \gamma, \sigma^2|X, y, \lambda, \tau_1, \tau_2, \kappa)\\
&\qquad \propto f(y|X, \beta, \gamma, \sigma^2) f(\beta|\sigma^2, \gamma, \lambda) f(\sigma^2|\tau_1, \tau_2) f(\gamma|\kappa).
\end{aligned}
\tag{3.2}
$$

With (3.2), we can estimate $(\beta, \gamma, \sigma^2)$ via various inference methods. In this paper, the *maximum a posteriori* (MAP) method is adopted. Formally, the MAP estimator for $(\beta, \gamma, \sigma^2)$ is defined by

$$
(\widehat{\beta}, \widehat{\gamma}, \widehat{\sigma}^2) = \arg\min_{\beta,\gamma,\sigma^2}\{-2\log f(\beta, \gamma, \sigma^2|X, y, \lambda, \tau_1, \tau_2, \kappa)\},
\tag{3.3}
$$

that is, the minimizer of the minus 2 logarithm of the joint posterior density. The minus 2 logarithm of the joint posterior density can be explicitly expressed as

$$
\begin{aligned}
-2\log f(\beta, \gamma, \sigma^2|X, y, \lambda, \tau_1, \tau_2, \kappa) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\gamma_j\beta_j\right)^2\\
&\quad + \frac{\lambda}{\sigma^2}\sum_{j=1}^{p}\gamma_j\beta_j^2\\
&\quad + \frac{2\tau_2}{\sigma^2} + (n + 2\tau_1 + 2)\log\sigma^2\\
&\quad + \sum_{j=1}^{p}\gamma_j\log\left\{\frac{2\pi\sigma^2(1-\kappa)^2}{\lambda\kappa^2}\right\} + \text{const.}
\end{aligned}
\tag{3.4}
$$

Here, we have used an equivalent representation $\text{Normal}(0, \sigma^2\lambda^{-1})^{\gamma_j} \times \mathbb{I}(\beta_j = 0)^{1-\gamma_j}$ for $f(\beta_j|\sigma^2, \lambda, \gamma)$ given that $\gamma_j \in \{0, 1\}$. Note that in (3.4) the term $\sum_{j=1}^{p} \log \mathbb{I}(\beta_j = 0)^{1-\gamma_j}$ vanishes since for every j, $\gamma_j = 1$ implies $\mathbb{I}(\beta_j = 0) = 0$. In turn, $(1 - \gamma_j) \log \mathbb{I}(\beta_j = 0) = 0 \cdot \infty = 0$. On the other hand, $\gamma_j = 0$ implies $\mathbb{I}(\beta_j = 0) = 1$, and in turn $\log \mathbb{I}(\beta_j = 0) = \log 1 = 0$.

For practical purposes, we fix σ^2 and multiply (3.4) with σ^2 in the following discussion. Given that σ^2 is fixed, the function (3.4) has some meaningful interpretations in terms of regularization estimation on β. For example, by definition $\gamma_j \geq 0$, and the quantity $\sum_{j=1}^{p} \gamma_j$ can be seen as an l_1 norm on the vector γ. Given the above argument, we can write the fourth term on the right-hand side of (3.4) as $\rho_{\lambda,\kappa,\sigma^2}\|\gamma\|_1$, where $\rho_{\lambda,\kappa,\sigma^2} = \sigma^2 \log[2\pi\sigma^2\lambda^{-1}(1 - \kappa)^2\kappa^{-2}]$. Note that as κ increases, $\rho_{\lambda,\kappa,\sigma^2}$ will decrease. It implies that a strong belief in the presence of a variable will decrease the penalty value for the variable. In addition, by definition $\gamma_j = \mathbb{I}(\beta_j \neq 0)$, and the term $\|\gamma\|_1$ can further be seen as an l_0 norm on β, as $\|\gamma\|_1 = \sum_{j=1}^{p} |\mathbb{I}(\beta_j \neq 0)| = \lim_{s\to 0} \sum_{j=1}^{p} \beta_j^s$, which is the l_0 norm by definition. Here, we have used the assumption that $0^0 = 0$. We can express the fourth term in (3.4) by $\rho_{\lambda,\kappa,\sigma^2}\|\beta\|_0$.

3.2. *Parameter estimation.* Now given all other parameters fixed, the MAP estimator of σ^2 can be derived by first making a derivative of (3.4) with respect to σ^2, setting the derivative to zero, and then solving the equation for σ^2. The estimation of β is further carried out given σ^2 is fixed. With fixed σ^2 and the interpretations of regularization estimation given above, (3.4) has an equivalent representation given by

$$L(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2}) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \rho_{\lambda,\kappa,\sigma^2}\|\beta\|_0 + \text{const}. \tag{3.5}$$

Note that here we have multiplied (3.4) with σ^2. Now with (3.5), we can construct an iteration scheme to obtain (3.3). At the $(m + 1)$th iteration, the iteration scheme is given by

$$\begin{aligned}
(\widehat{\sigma}^2)^{(m+1)} &= \frac{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\widehat{\gamma}_j^{(m)}\widehat{\beta}_j^{(m)})^2 + \lambda\sum_{j=1}^{p}\widehat{\gamma}_j^{(m)}(\widehat{\beta}_j^{(m)})^2 + 2\tau_2}{n + \sum_{j=1}^{p}\widehat{\gamma}_j^{(m)} + 2\tau_1 + 2},\\
\widehat{\beta}^{(m+1)} &= \arg\min_{\beta} L\big(\beta; \lambda, \rho_{\lambda,\kappa,(\widehat{\sigma}^2)^{(m+1)}}\big),\\
\widehat{\gamma}^{(m+1)} &= \big(\mathbb{I}(\widehat{\beta}_1^{(m+1)} \neq 0), \mathbb{I}(\widehat{\beta}_2^{(m+1)} \neq 0), \ldots, \mathbb{I}(\widehat{\beta}_p^{(m+1)} \neq 0)\big).
\end{aligned} \tag{3.6}$$

Note that the objective function (3.5) involves an l_0 norm, which by definition, is not continuous. Therefore, related optimization tasks in the second term of (3.6) require some refinements. Here, we adopt a relaxation approach to tackling the optimization problem. We begins the approach by noting that, mathematically the l_0

norm on a p-dimensional vector β can be expressed as

$$\|\beta\|_0 = \lim_{\tau_3 \to 0} \sum_{j=1}^{p} \frac{\log(1+\tau_3^{-1}|\beta_j|)}{\log(1+\tau_3^{-1})}, \tag{3.7}$$

which can be verified by seeing (3.7) as a function of τ_3 and using l'Hôpital's rule. A more detailed discussion on the properties of the log-sum function on the right-hand side of (3.7) is given in Supplementary Material [35]. With representation (3.7), the objective function (3.5) can be reexpressed as

$$\begin{aligned} L(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2}) &= \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \\ &\quad + \rho_{\lambda,\kappa,\sigma^2}\left\{\lim_{\tau_3 \to 0} \sum_{j=1}^{p} \frac{\log(1+\tau_3^{-1}|\beta_j|)}{\log(1+\tau_3^{-1})}\right\} + \text{const.} \end{aligned} \tag{3.8}$$

If τ_3 is small enough, the log-sum function on the right-hand side of (3.8) will give an approximate representation of $\|\beta\|_0$. Graphical representations for the log-sum function with different τ_3 and their mixtures with the squared l_2 norm can be found in the left and middle panels of Figure 1. In addition, since the log-sum function in (3.8) is continuous in β, the combinatorial nature of $\|\beta\|_0$ is relaxed. However, the term $\log(1+\tau_3^{-1}|\beta_j|)$ is not convex in β_j, and replacing $\|\beta\|_0$ with (3.7) in (3.5) still makes objective function (3.8) remain nonconvex. To tackle this problem, a majorization–minimization algorithm is adopted. Majorization-minimization (MM) algorithms [18, 34] are a set of analytic procedures aiming to tackle difficult optimization problems by modifying their objective functions so that solution spaces of the modified ones are easier to explore. For an objective function $g(\theta)$, the modification procedure relies on finding a function $h(\theta; \theta^{(l)})$ satisfying the following properties:

$$\begin{aligned} h(\theta; \theta^{(l)}) &\geq g(\theta) \qquad \text{for all } \theta, \\ h(\theta^{(l)}; \theta^{(l)}) &= g(\theta^{(l)}). \end{aligned} \tag{3.9}$$

In (3.9), the objective function $g(\theta)$ is said to be majorized by $h(\theta; \theta^{(l)})$. In this sense, $h(\theta; \theta^{(l)})$ is called the majorization function. In addition, (3.9) implies that $h(\theta; \theta^{(l)})$ is tangent to $g(\theta)$ at $\theta^{(l)}$. Moreover, if $\theta^{(l+1)}$ is a minimizer of $h(\theta; \theta^{(l)})$, then (3.9) further implies that

$$g(\theta^{(l)}) = h(\theta^{(l)}; \theta^{(l)}) \geq h(\theta^{(l+1)}; \theta^{(l)}) \geq g(\theta^{(l+1)}), \tag{3.10}$$

which means that the iteration procedure $\theta^{(l)}$ pushes $g(\theta)$ toward its minimum.

Now we turn back to the function on the right-hand side of (3.8). Note that, since $\log(\theta)$ is a concave function of θ for $\theta > 0$, therefore the inequality

$$\log(\theta') + \frac{\theta}{\theta'} - 1 \geq \log(\theta) \tag{3.11}$$

holds for all $\theta > 0$ and $\theta' > 0$. Note that the left-hand side of (3.11) is convex in θ. In addition, if we let $\theta' = \theta$, then (3.11) becomes an equality, which implies that the left-hand side of (3.11) satisfies the properties stated in (3.9), therefore is a valid function for majorizing $\log(\theta)$.

PROPOSITION 3.1. *Define $\rho_{\tau_3} = 1/\log(1 + \tau_3^{-1})$ and let $L'(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2})$ be the same as (3.5) but without the constant term. Then $L'(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2})$ can be majorized by the following function*:

$$L''(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2}, \beta') = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \rho_{\lambda,\kappa,\sigma^2} h_2(\beta; \beta'), \tag{3.12}$$

where

$$h_2(\beta; \beta') = \lim_{\tau_3 \to 0} \rho_{\tau_3} \sum_{j=1}^{p} \biggl(\log(1 + \tau_3^{-1}|\beta'_j|) + \frac{|\beta_j| + \tau_3}{|\beta'_j| + \tau_3} - 1\biggr). \tag{3.13}$$

PROOF. Let $h_1(\beta) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$. Assume $\beta^{(l+1)}$ minimizes $L''(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2}, \beta')$ given $\beta' = \beta^{(l)}$. Then with (3.7) and the inequality (3.11), the quantity $L'(\beta^{(l+1)}; \lambda, \rho_{\lambda,\kappa,\sigma^2})$ can be bounded in a way such that

$$\begin{aligned} L'(\beta^{(l+1)}; \lambda, \rho_{\lambda,\kappa,\sigma^2}) &= h_1(\beta^{(l+1)}) + \rho_{\lambda,\kappa,\sigma^2} \lim_{\tau_3 \to 0} \rho_{\tau_3} \sum_{j=1}^{p} \log(1 + \tau_3^{-1}|\beta_j^{(l+1)}|) \\ &\le h_1(\beta^{(l+1)}) + \rho_{\lambda,\kappa,\sigma^2} h_2(\beta^{(l+1)}; \beta^{(l)}) \qquad (3.14) \\ &= L''(\beta^{(l+1)}; \lambda, \rho_{\lambda,\kappa,\sigma^2}, \beta^{(l)}) \end{aligned}$$

which verifies the first condition stated in (3.9). For $\beta = \beta'$, $h_2(\beta; \beta')$ is equal to the log-sum function in (3.7), which verifies the second condition stated in (3.9) and completes the proof. □

A graphical representation of using MM algorithms in approximating the log-sum function in (3.7) can be found in the right panel of Figure 1. From the argument given above, we can construct an iteration scheme to obtain the minimizer of $L(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2})$, with the l_0 norm, or equivalently the log-sum function, replaced by $h_2(\beta; \beta')$ defined in Proposition 3.1. For example, in (3.6), $\widehat{\beta}^{(m+1)}$ can be obtained by carrying out the following iteration scheme:

$$\begin{aligned} &\widehat{\beta}^{(m+1,l+1)} \\ &\quad = \arg\min_{\beta} \Biggl\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \rho_{\lambda,\kappa,(\widehat{\sigma}^2)^{(m+1)}} \sum_{j=1}^{p} \widehat{\phi}_j^{(m+1,l)} |\beta_j| \Biggr\} \end{aligned} \tag{3.15}$$

over index l, where $\widehat{\phi}_j^{(m+1,l)} = \lim_{\tau_3 \to 0}[\log(1 + \tau_3^{-1})(|\widehat{\beta}_j^{(m+1,l)}| + \tau_3)]^{-1}$. The procedure of using iteration scheme (3.15) in obtaining the minimizer for the objective

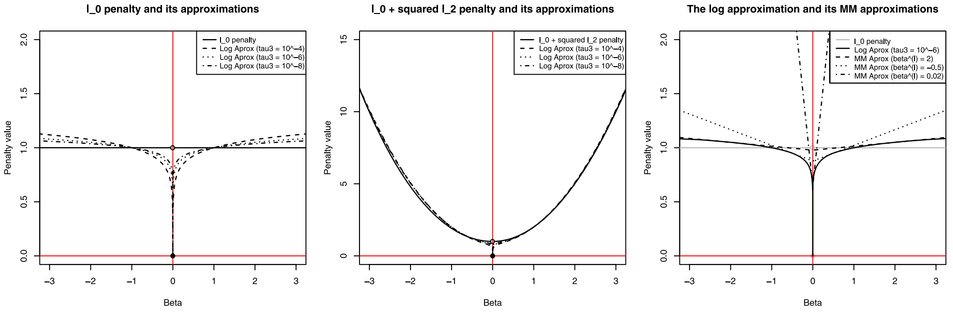

FIG. 1. *The penalty functions and related approximations.*

function (3.5), or equivalent (3.8), is called the BAVA-MIO (BAyesian VAriable selection using a Majorization–mInimization apprOach), and the resulting minimizer is called the BAVA-MIO estimator.

Note that the last term on the right-hand side of (3.15) is a linear combination of $\widehat{\phi}_j^{(m+1,l)}|\beta_j|$, a convex function of β_j, therefore given $\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$ is convex in β, the whole objective function in (3.15) will be convex in β, which guarantees that the iteration scheme will converge. In addition, the minimizer (3.15) can be obtained by using the coordinate descent algorithm proposed by Friedman et al. [9]. In practice, the coordinate descent algorithm is based on iteratively cycling a one-dimensional soft-thresholding scheme. Given that $\rho_{\lambda,\kappa,\sigma^2}$ is fixed, at the $(m_1 + 1)$th iteration, the soft-thresholding scheme for the jth coordinate is given by

$$(3.16) \qquad \tilde{\beta}_j^{(m_1+1)} = \left(\sum_{i=1}^n x_{ij}^2 + \lambda\right)^{-1} \mathrm{ST}\left(\sum_{i=1}^n x_{ij}\tilde{r}_{i,-j}^{(m_1)}, \rho_{\lambda,\kappa,\sigma^2}\frac{\tilde{\phi}_j^{(m_1)}}{2}\right),$$

where $\tilde{r}_{i,-j}^{(m_1)} = y_i - \sum_{j' \neq j} x_{ij'}\tilde{\beta}_{j'}^{(m_1')}$, with $m_1' = m_1 + 1$ for $j' = 1, 2, \dots, j-1$, and $m_1' = m_1$ for $j' = j+1, j+2, \dots, p$, and $\tilde{\phi}_j^{(m_1)} = \lim_{\tau_3 \to 0}[\log(1+\tau_3^{-1})(|\tilde{\beta}_j^{(m_1)}| + \tau_3)]^{-1}$. Here $\mathrm{ST}(a, b)$ is a soft-thresholding operator defined by $\mathrm{ST}(a, b) = \mathrm{sign}(a)(|a| - b)_+$. A detailed derivation of (3.16) is given in Appendix A of Supplementary Material [35].

3.3. *Choosing hyperparameters.* Choosing appropriate hyperparameters for prior construction is an important issue in many Bayesian inference problems. For hyperparameters present in the model (3.1), we consider the triple $(\lambda, \tau_1, \tau_2)$ first. One principle we adopt in parameterizing the hyperparameters is that as the number of samples n increases, the impact of the hyperparameters in parameter estimation will become less significant. In addition, we let $\tau_1 = \tau_2 + 1$, so that the prior expectation of σ^2 is equal to 1. Given these conditions, one of the possible choices is $(\lambda, \tau_1, \tau_2) = (1/\sqrt{n}, p\log p/\sqrt{n} + 1, p\log p/\sqrt{n})$. We will discuss other possible settings in the simulation study in the later section.

Now we consider the prior inclusion probability κ. In some circumstances, data-driven empirical Bayes approaches [12] are proposed to obtain κ, while in other circumstances full Bayesian methods that assign priors on κ are proposed. For example, please see [24]. Unlike previously proposed approaches, in which single point estimates were obtained for κ, we adopt an approach by specifying a feasible region for the function

$$\psi(\kappa) = \tfrac{1}{2}\left[\sigma^2 \log\left(2\pi\sigma^2\lambda^{-1}(1-\kappa)^2/\kappa^2\right)\right]\widehat{\phi}^{(0)}, \tag{3.17}$$

and carry out parameter estimation under different values of $\psi(\kappa)$. Here we have assumed $\widehat{\phi}_j^{(0)} = \widehat{\phi}^{(0)}$ for $j = 1, 2, \dots, p$. Note that by definition the term $\widehat{\phi}_j^{(0)}$ is a function of the initial value $\widehat{\beta}_j^{(0)}$. The function $\psi(\kappa)$ is the threshold used in the soft-thresholding scheme (3.16). We carry out the parameter estimation with values in the feasible region and look for which values of $\psi(\kappa)$ lead to the best performance measured by criteria such as ten fold cross validation or the Bayes factor. Under this approach, estimated parameters can be seen as functions of κ on the feasible region. Given different values of κ, curve-like paths for estimated parameters can be obtained. The main reason we adopt this "whole-path" fitting strategy is that the optimization procedure may get stuck in some stationary points. It can occur in a situation in which we need an initial value $\widehat{\phi}_j^{(1)}$ to run the iteration scheme (3.15). By definition, $\widehat{\phi}_j^{(1)}$ is a function of $\widehat{\beta}_j^{(1)}$, which by definition, is a function of $\psi(\kappa)$. As pointed out by Candés et al. [3] and Mazumder et al. [25], different $\widehat{\phi}_j^{(1)}$ may lead to different solutions for the minimizer. Under this situation, a global minimum may not be guaranteed. By using the strategy given above, we can run the iteration scheme (3.15) with a large number of possible values of $\widehat{\phi}_j^{(1)}$, therefore eliminating the possibility that the solution is stuck in some local minima.

Our approach is similar to the one using a fixed grid on the tuning parameter and then running parameter estimation with different values of the tuning parameter. This fixed grid approach to tuning parameter selection has been adopted in [9, 11, 27] and is advocated by [10, 34] for fast and accurate parameter estimation.

3.4. *A toy example.* Here, we provide a toy example to illustrate the BAVA-MIO estimation. We let the number of samples $n = 100$ and the number of covariates $p = 1{,}000$. For regression coefficients $\beta = (\beta_1, \beta_2, \dots, \beta_{1{,}000})$, we let $\beta_{250} = 2$, $\beta_{500} = -3.2$, $\beta_{750} = -1.25$, $\beta_{1{,}000} = 5.44$, and $\beta_j = 0$ for all j's $\in \{1, 2, \dots, 1{,}000\} \setminus \{250, 500, 750, 1{,}000\}$. We generate each row of X independently identically from $\mathrm{MVN}(0, I_{p\times p})$, and then calculating $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathrm{MVN}(0, I_{n\times n})$. For the hyperparameters, we let $\tau_1 = 0.2p\log(p)/\sqrt{n} + 1$, $\tau_2 = 0.2p\log(p)/\sqrt{n}$ and $\lambda = 1/\sqrt{n}$. Further let $\tau_3 = 10^{-6}$. We use 100 equal spaced points to form a grid for $\Psi(\kappa)$. We perform two BAVA-MIO estimations: one uses the Bayes factor and the other uses ten fold cross validation for tuning parameter selection. Remember the index set S is defined by $S = \{j : \gamma_j = 1\}$. We define the Bayes factor between models $\mathcal{M}_{S'}$ and $\mathcal{M}_S$ by

$$\text{BF}(\mathcal{M}_{S'}, \mathcal{M}_S; y) = \frac{f(y|\gamma', \tau_1, \tau_2, \kappa, \lambda)}{f(y|\gamma, \tau_1, \tau_2, \kappa, \lambda)}, \tag{3.18}$$

where the term $f(y|\gamma, \tau_1, \tau_2, \kappa, \lambda)$ refers to the marginalized likelihood with β and σ^2 being integrated out with respect to their prior probability measures. For the Bayesian model stated in (3.1), the marginalized likelihood has a closed form representation given by

$$\begin{aligned} f(y|\gamma, \tau_1, \tau_2, \kappa, \lambda) = {} & \frac{\pi^{-n/2}}{|\lambda^{-1} X_S^T X_S + I_\gamma|^{1/2}} \frac{(2\tau_2)^{\tau_1}}{\Gamma(\tau_1)} \Gamma\left(\frac{n + 2\tau_1}{2}\right) \\ & \times \left(y^T (\lambda^{-1} X_S X_S^T + I_n)^{-1} y + 2\tau_2\right)^{-[(n+2\tau_1)/2]}. \end{aligned}$$

In subsequent sections we will use the measure (3.18) for variable selection. In addition, for all variable selection tasks using (3.18), the baseline model $\mathcal{M}_S$ will always refer to the null model.

The results are shown in Figure 2. The path plot in the top left panel of Figure 2 shows that nonzero coefficients entered into the model earlier under the BAVA-MIO estimation. In addition, the paths of estimated coefficients behave similar to those under the hard-thresholding estimation, that is, once a coefficient is estimated to be nonzero, the corresponding estimation path makes a sharp jump to the nonthresholded value. Moreover, due to the presence of the squared l_2 norm in the objective function, the number of selected covariates can be larger than the number of samples. Throughout the estimation procedure, the maximum number of selected covariates is 831, which is much larger than the number of samples $n = 100$. Here we also provide the lasso estimation for regression fitting with the same data. The results are shown in the bottom panel of Figure 2. As compared with the lasso estimation, in which 33 covariates are selected using ten fold cross validation, the BAVA-MIO estimations using the Bayes factor and ten fold cross validation correctly select covariates with nonzero coefficients. In addition, as shown in the right panel of Figure 2, values of the nonzero coefficients are also estimated more accurately under the BAVA-MIO estimations.

4. Simulation studies.

In this section, we conduct two simulation studies. The first one is a general assessment on the performance of the BAVA-MIO estimation. The second one focuses the performance of the BAVA-MIO estimation under various situations in which the Irrepresentable Condition may or may not hold.

4.1. *Simulation study I.*

In the first simulation study, we compare the BAVA-MIO estimation with other estimation approaches by fitting regression model $Y = X\beta + \varepsilon$ with data generated from different simulation schemes. Here Y and ε

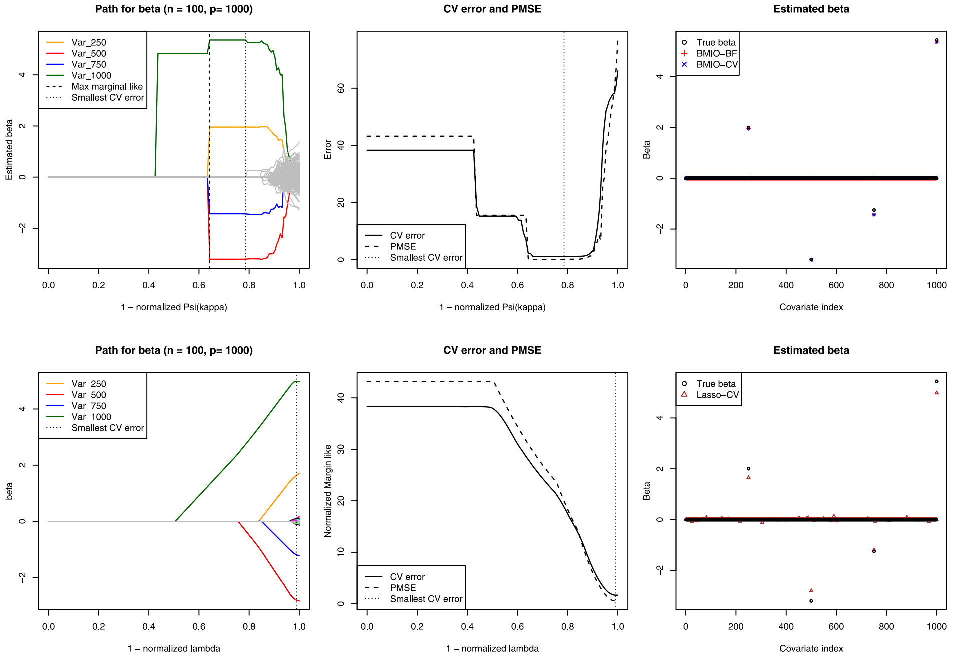

FIG. 2. *Estimation results for the toy example.*

are n-dimensional vectors, X is an $n \times p$ matrix and β is a p-dimensional vector. We assume each entry in ε is i.i.d. from Normal$(0, \sigma_Y^2)$, and each row in the design matrix X is i.i.d. from MVN$(0, \Sigma_X)$. Throughout the whole simulation study, we let $p = 120$. For regression coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, we generate β_j from Normal$(0, 1)$ for $j = 1, 2, \ldots, 10$ and let $\beta_j = 0$ for $j = 11, 12, \ldots, 120$. That is, we have 10 nonzero and 110 zero coefficients in the "true" model. In addition, we use different values of $(\Sigma_X, \sigma_Y^2, n)$ in generating the design matrix X and the error term ε. We apply three different Σ_X to generate the design matrix. The first one has an independent structure with diagonal terms equal to 1 and off-diagonal terms equal to 0. The second one has a covariance structure such that $(\Sigma_X)_{ij} = 1$ for $i = j$ and $(\Sigma_X)_{ij} = 0.5$ for $i \neq j$. The third one has a covariance structure such that $(\Sigma_X)_{ij} = 0.5^{|i-j|}$. Now define the signal-to-noise ratio by SNR $= \sqrt{\mathbb{E}(\beta^T \Sigma_X \beta)/\sigma_Y^2}$. We consider $\sigma_Y^2 = 10, 1$ and 0.2 in generating the error term ε. For $\Sigma_X = I_{10\times 10}$, these values correspond to SNR $= 1, 3.16$ and 7.07, respectively. For practical purposes, we will use the labels SNR $= 1$ for experiments using $\sigma_Y^2 = 10$, SNR $= 3.16$ for experiments using $\sigma_Y^2 = 1$ and SNR $= 7.07$ for experiments using $\sigma_Y^2 = 0.2$. By using X, β and ε, the response vector Y is calculated by $Y = X\beta + \varepsilon$. For the number of samples, we consider five values $n = 40, 80, 120, 160$ and 200. With three different structures for Σ_X, three different values for σ_Y^2, and five different values for n, we have total $3 \times 3 \times 5 = 45$ simulation experiments.

Here we describe hyperparameter settings in BAVA-MIO estimations. We let hyperparameters $(\lambda, \tau_1, \tau_2) = (1/\sqrt{n}, p \log p/\sqrt{n} + 1, p \log p/\sqrt{n})$ for the cases of SNR $= 3.16$ and 7.07. For the case of SNR $= 1$, we use

$$\begin{aligned}
&\lambda = \left(\frac{1 - \widehat{\mathrm{corr}}}{\widehat{\mathrm{corr}}}\right)^2 \sqrt{\frac{p}{n}} \log p, \\
(4.1) \qquad &\tau_1 = \left(\frac{\widehat{\mathrm{corr}}}{1 - \widehat{\mathrm{corr}}}\right)\left(\frac{p \log p}{\sqrt{n}}\right)^{\widehat{\mathrm{corr}}/\log n} + 1, \\
&\tau_2 = \frac{1}{\sqrt{n}}\left(\frac{p \log p}{\sqrt{n}}\right)^{1+\widehat{\mathrm{corr}}/\log n},
\end{aligned}$$

where $\widehat{\mathrm{corr}}$ is an average over the top 10 percent absolute values of the sample correlations between response Y and covariates X. For tuning parameter selection, we use two criteria: the Bayes factor, which is defined in (3.18), and ten-fold cross validation. The resulting estimators are called BMIO-BF and BMIO-CV, respectively.

We also carry out three other estimation approaches for comparisons. The first one is the lasso [33]. We use R package "glmnet" to obtain the lasso estimates. The tuning parameter is selected using ten fold cross validation. The second approach is the relaxed lasso [28]. We use R package "relaxo," which is the companion software to [28], to obtain the relaxed lasso estimates. The tuning parameter is selected using ten fold cross validation with 100 values of scaling parameters equally spaced in [0, 1]. The third approach is the adaptive lasso [40]. We use R package "parcor" to obtain the adaptive lasso estimates with the default setting that uses the lasso estimate as the initial value for the weight and selects the tuning parameter λ via ten fold cross validation.

We collect several performance measures at each simulation run. The first one is the standardized l_2 distance between a given estimate $\widehat{\beta}$ and the true regression coefficient vector β, which is defined by

$$l_2\text{-dis}(\widehat{\beta}) = \sqrt{\frac{\sum_{j=1}^p (\widehat{\beta}_j - \beta_j)^2}{\sum_{j=1}^p \beta_j^2}}.$$

The second one is the predictive mean squared error of $\widehat{\beta}$ for a test data set, which is defined by

$$\mathrm{PMSE}(\widehat{\beta}) = \frac{\sum_{i=1}^{n_{\mathrm{test}}} (x_{i,\mathrm{test}}^T \widehat{\beta} - x_{i,\mathrm{test}}^T \beta)^2}{n_{\mathrm{test}}}.$$

The test data set contains $n_{\mathrm{test}} = n \times 10$ data points generated using a simulation scheme the same as the training data set. The third one is the number of coefficients with nonzero estimated values $|\widehat{S}|$, where $\widehat{S} = \{j : \widehat{\beta}_j \neq 0\}$. The final one is the sign function-based false positive rates, which is defined by

$$\text{S-FPR} = \frac{\#\{j \in \widehat{S} : \operatorname{sign}(\widehat{\beta}_j) \neq \operatorname{sign}(\beta_{\text{true},j})\}}{|\widehat{S}|},$$

where the sign function sign$(\cdot)$ is defined in Section 2.

For each of the 45 simulation experiments, we generate 100 runs to collect the four performance measures. We then plot the average of each performance measure against the ratio $n/|S|$, that is, the ratio between the number of samples and the number of true coefficients with nonzero values. These plots are shown in Figures 3, 4 and 5 for SNR $= 1, 3.16$ and 7.07, respectively. From the three figures, we can see none of the estimation approaches can dominate the others in all four performance measures. In most cases, BAVA-MIO based estimations have smaller sign function-based false positive rates, as shown in the second column of

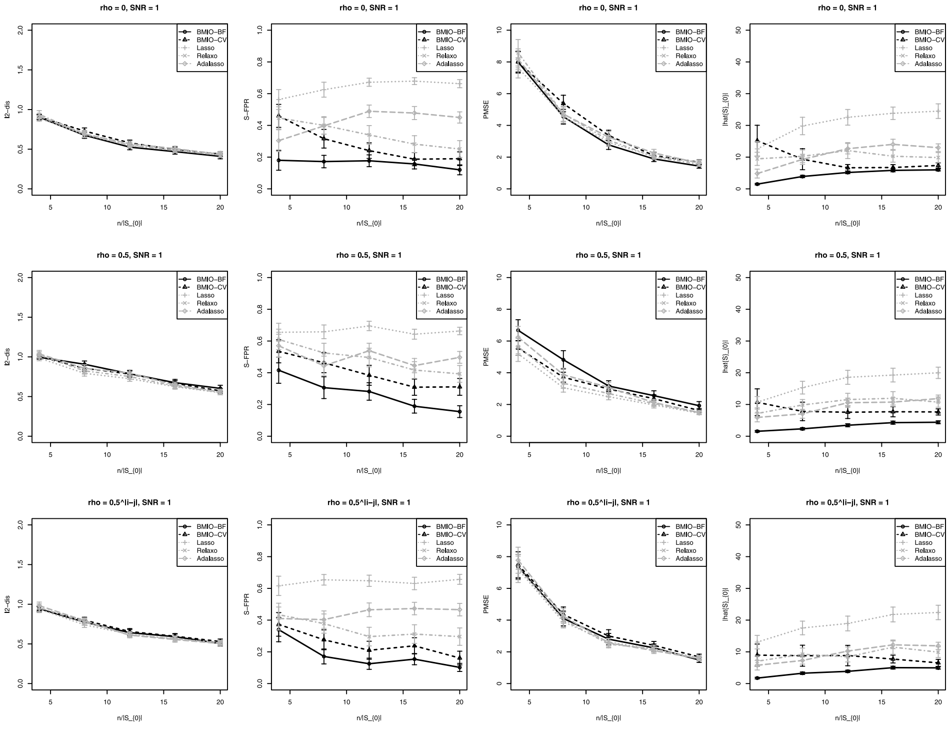

FIG. 3. *Simulation results given SNR* = 1. *Top*: *Model* 1 (*covariance matrix with off-diagonal terms equal to* 0); *Middle*: *Model* 2 (*covariance matrix with off-diagonal terms equal to* 0.5); *Bottom*: *Model* 3 (*covariance matrix with off-diagonal terms following a specified covariance structure*). *First column*: *standardized* l_2-*distance between estimated and true values*; *Second column*: *sign function-adjusted false positive rate*; *Third column*: *prediction mean squared error*; *Fourth column*: *number of nonzero estimates*.

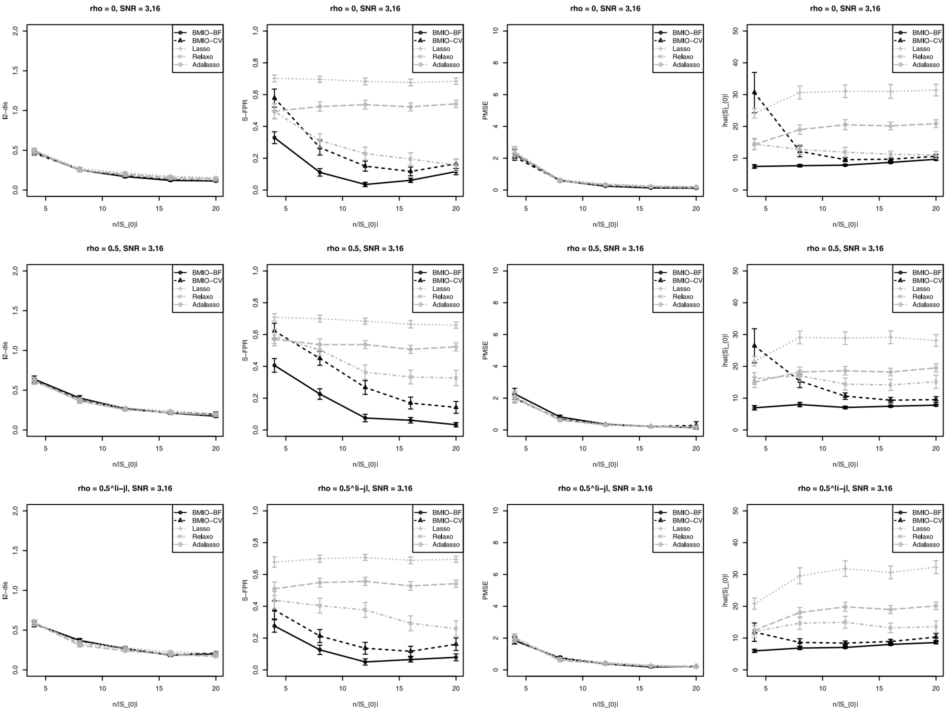

FIG. 4. *Simulation results given SNR* $= 3.16$. *Top*: *Model* 1 (*covariance matrix with off-diagonal terms equal to* 0); *Middle*: *Model* 2 (*covariance matrix with off-diagonal terms equal to* 0.5); *Bottom*: *Model* 3 (*covariance matrix with off-diagonal terms following a specified covariance structure*). *First column*: *standardized l_2-distance between estimated and true values*; *Second column*: *sign function-adjusted false positive rate*; *Third column*: *prediction mean squared error*; *Fourth column*: *number of nonzero estimates*.

each figure. It implies that more accurate variable selection may be done using the BAVA-MIO estimations. These findings become more significant as the number of samples increases. In addition, BAVA-MIO estimations have fewer numbers of nonzero estimates, as shown in the fourth column of each figure. Moreover, since the BAVA-MIO estimation using the Bayes factor has relatively fewer numbers of nonzero estimates, it is surprising that the PMSE and l_2-dis measures under the BMIO-BF estimation are comparable to those under other estimation approaches, for example, in the cases with SNR $= 3.16$ and in some cases with SNR $= 1$. However, we also noticed that the BMIO-BF estimation has higher values in the PMSE and l_2-dis in the cases with SNR $= 7.07$, particularly in the situations in which the number of samples is small.

4.2. *Simulation study II.* In the second simulation study, we investigate the impact of the Irrepresentable Condition on the performance of BAVA-MIO estimation in variable selection. Before stating the Irrepresentable Condition, we give

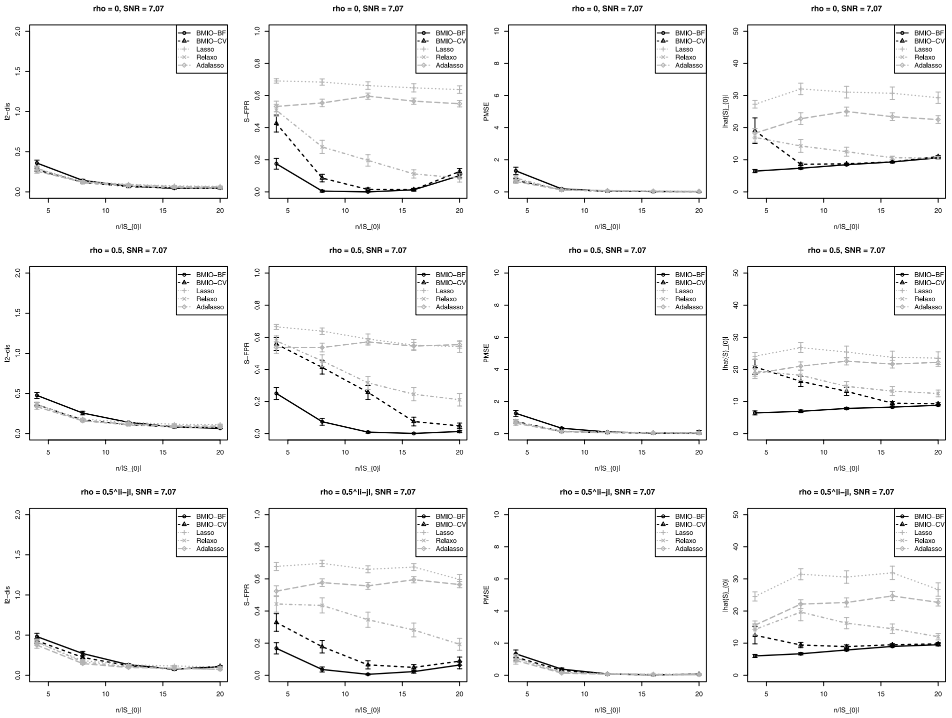

FIG. 5. *Simulation results given SNR* = 7.07. *Top*: *Model* 1 (*covariance matrix with off-diagonal terms equal to* 0); *Middle*: *Model* 2 (*covariance matrix with off-diagonal terms equal to* 0.5); *Bottom*: *Model* 3 (*covariance matrix with off-diagonal terms following a specified covariance structure*). *First column*: *standardized l_2-distance between estimated and true values*; *Second column*: *sign function-adjusted false positive rate*; *Third column*: *prediction mean squared error*; *Fourth column*: *number of nonzero estimates*.

some notation definitions. We define $S_0 = \{j : \beta_j \neq 0, \text{ for some } j \in \{1, 2, \dots, p\}$ and $S_0^c = \{1, 2, \dots, p\} \setminus S_0$. Let β_{S_0} denote the coefficients with indices in S_0 and $\beta_{S_0^c}$ the coefficients with indices in S_0^c. Similar definitions are also applied to X_{S_0} and $X_{S_0^c}$. An estimator $\widehat{\beta}(n)$ is said to be sign consistent in estimating β if the probability of the event $\{\text{sign}(\widehat{\beta}(n)) = \text{sign}(\beta)\}$ approaches to 1 as $n \to \infty$. Given the sign consistency holds, the estimated index set $\widehat{S}_0 = \{j : \widehat{\beta}_j \neq 0\}$ will be the same as the true index set S_0, therefore the sign consistency implies variable selection consistency, that is, asymptotically with probability one, nonzero-valued coefficients will have nonzero estimated values and zero-valued coefficients will be estimated with zero values.

Zhao and Yu [39] showed that if one wants the lasso estimation to achieve the sign consistency, then the design matrices X must satisfy the following condition:

$$\|X_{S_0^c}^T X_{S_0} (X_{S_0}^T X_{S_0})^{-1} \text{sign}(\beta_{S_0})\|_\infty < 1, \tag{4.2}$$

where β_{S_0} is the vector of nonzero-valued coefficients. The condition (4.2) is called the (Weak) Irrepresentable Condition. If the Irrepresentable Condition (4.2) fails to hold, then the sign consistency will never occur even when $n \to \infty$. An intuitive way to explain the Irrepresentable Condition is to see the quantity $X^T_{S_0^c} X_{S_0}(X^T_{S_0} X_{S_0})^{-1}$ as a least squares estimate for the regression $X_{S_0^c}$ on X_{S_0}. In this sense, the Irrepresentable Condition states that the largest amount of coefficients for the regression $X_{S_0^c}$ on X_{S_0} should not exceed 1, that is, $X_{S_0^c}$ is "irrepresentable" in terms of X_{S_0}.

Here we conduct a simulation study for the investigation. We generate 100 design matrices in which each row is i.i.d. from $\mathrm{MVN}(0, \Sigma_X)$, with $\Sigma_X \sim \mathrm{Wishart}(I_{p\times p}, p, p)$ with $p = 30$. This setting is similar to the one used in Zhao and Yu's study. Corresponding regression coefficients β are generated in a way that the first 5 entries of β are i.i.d. from Normal(0, 1), and the rest of 25 entries are set to 0. Note that for some pairs (X, β), the Irrepresentable Condition (4.2) will hold, but for some pairs it will not hold. Zhao and Yu defined the irrepresentable statistic by

$$\text{Irr.stat} = 1 - \|X^T_{S_0^c} X_{S_0}(X^T_{S_0} X_{S_0})^{-1} \operatorname{sign}(\beta_{S_0})\|_\infty. \tag{4.3}$$

The Irrepresentable Condition is considered to be violated if Irr.stat is smaller than zero. We carry out 100 simulation runs and calculate the irrepresentable statistic for each pair (X, β). In each run, we generate $n = 100$ data points. We use $\sigma_Y^2 = 0.05$ to generate the error term ε, which is corresponding to SNR $= 10$. We then fit the regression model using the lasso estimation and the BAVA-MIO estimation with these data points. For each pair (X, β), we calculate the model selection probability $P(\widehat{S}_0 = S_0)$ based on counting the times of whether the estimated sign vector matched the true sign vector throughout the whole regularization paths.

We also carry out the same simulation experiment under SNR $= 5$, 2 and 1. The scatter plots in Figure 6 show the estimated model selection probability against Irr.stat under signal-to-noise ratios SNR $= 10$, 5, and 2. From these plots, we can

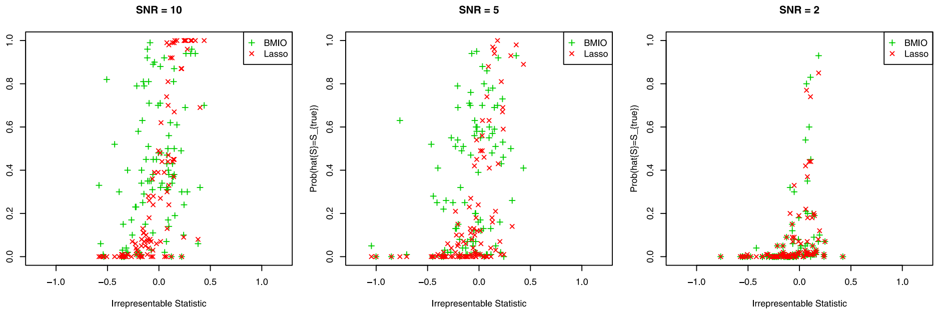

FIG. 6. *Scatter plots for the sign probability $\mathbb{P}(\widehat{S} = S_0)$ against the Irrepresentable Statistic under different signal-to-noise ratios.*

TABLE 1

The sign probability $\mathbb{P}(\widehat{S} = S_0)$ under different signal-to-noise ratios. Each value is calculated by averaging over 100 *simulation runs, and the corresponding standard error is given in the bracket. The term corr. in the second line of each panel is the squared correlation between the sign probability and the irrepresentable statistic. We use Kendall's τ for the correlation calculation*

Name	**SNR = 10**	**SNR = 5**	**SNR = 2**	**SNR = 1**
BMIO	0.398 (0.030)	0.338 (0.030)	0.077 (0.018)	0.023 (0.006)
corr.	0.052	0.047	0.180	0.110
Lasso	0.314 (0.047)	0.203 (0.030)	0.072 (0.016)	0.022 (0.006)
corr.	0.418	0.232	0.194	0.107

see that performances of the BAVA-MIO and the lasso estimations are deteriorated when the signal-to-noise ratio is decreasing. However, we also found in some circumstances the BAVA-MIO estimation can achieve high model selection probabilities even when the Irrepresentable Condition is violated, that is, Irr.stat is smaller than zero. In Section 5, we will provide a theoretical result to explain this phenomenon. The second and fourth rows in Table 1 show the squared correlations between the estimated model selection probability and the irrepresentable statistic. The squared correlations for the BAVA-MIO estimation are relatively small in comparison with the lasso estimation.

5. Asymptotic analysis.

In this section, we will derive asymptotic results for the BAVA-MIO estimator. When deriving the asymptotic results, we will consider a situation in which the number of parameters p is an increasing function of the number of samples n. For practical purposes, we will focus on the case $p = p(n) \propto n^{\alpha}$, where $\alpha > 0$. The first asymptotic result gives a theoretical explanation for the invariance of the BAVA-MIO estimator under the Irrepresentable Condition. The second result is on the posterior model consistency related to the hierarchical Bayesian formulation (3.1), and the third result shows the estimation consistency of the BAVA-MIO estimator.

5.1. *Sign consistency.*

Before stating the result of sign consistency, we give some notation definitions first. We use the same definitions given in Section 4.2 for S_0, S_0^c, β_{S_0}, $\beta_{S_0^c}$, X_{S_0} and $X_{S_0^c}$. Further let $\mathcal{S}$ denote the space that S_0 belongs to. For a symmetric matrix C, let $\Lambda_{\min}(C)$ and $\Lambda_{\max}(C)$ denote the smallest and the largest eigenvalues, respectively.

Our result on the sign consistency of the BAVA-MIO estimator is based on the following simplification: the variable σ^2 is fixed and the term $\rho_{\lambda,\kappa,\sigma^2}$ in (3.5) is treated as a constant. For simplicity, we let $\rho = \rho_{\lambda,\kappa,\sigma^2}$. Now define

$$\widehat{\beta}^{\tau_3} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \rho \sum_{j=1}^{p} \frac{\log(1 + \tau_3^{-1}|\beta_j|)}{\log(1 + \tau_3^{-1})}. \tag{5.1}$$

Note that the log-sum function on the right-hand side of (5.1) becomes $\|\beta\|_0$ if $\tau_3 \to 0$, and $\widehat{\beta}^{\tau_3}$ in this sense can be seen as the BAVA-MIO estimator. Now define

$$E_{0,\tau_3} = \{\beta : \operatorname{sign}(\beta_j) = \operatorname{sign}(\widehat{\beta}_j^{\tau_3}) \text{ for } j = 1, 2, \dots, p\}, \tag{5.2}$$

that is, the event of sign consistency for the estimator $\widehat{\beta}^{\tau_3}$ in (5.1). For practical purposes, further define $C_{SS_0} = n^{-1}(X_{S_0}^T X_{S_0} + \lambda I)$, $C_{S^c S_0} = n^{-1} X_{S_0^c}^T X_{S_0}$, $D_{S_0} = n^{-1/2} X_{S_0}^T \varepsilon$ and $D_{S_0^c} = n^{-1/2} X_{S_0^c}^T \varepsilon$. In the following, we give some assumptions that will be used in deriving the asymptotic results.

ASSUMPTION 1. For $C_{SS} = n^{-1}(X_S^T X_S + \lambda I)$ and any $S \in \mathcal{S}$, the maximum eigenvalue $\Lambda_{\max}(C_{SS})$ and the minimum eigenvalue $\Lambda_{\min}(C_{SS})$ satisfy the following condition:

$$0 \le c_1 < \Lambda_{\min}(C_{SS}) \le \Lambda_{\max}(C_{SS}) \le c_2 < \infty.$$

ASSUMPTION 2. For the vector $X^T\varepsilon$, $\|X^T\varepsilon\|_1 = O(p)$.

ASSUMPTION 3. For parameter λ, we assume $0 \le \lambda < \infty$. For parameter ρ, we assume $0 \le \rho$ and $\rho n^{-1/2} \to 0$.

Assumption 1 is a special case of the Restricted Eigenvalue Assumption stated in Bickel et al. [1]. It implies that the inverse of C_{SS_0} exists and the ratio $[\Lambda_{\max}(X_S^T X_S) + \lambda]/[\Lambda_{\min}(X_S^T X_S) + \lambda] \le c_2/c_1$ is bounded from above for any $S \in \mathcal{S}$. Assumption 2 is equivalent to the statement that $n^{-1/2}\|X^T\varepsilon\|_1$ is bounded from some quantity proportional to $pn^{-1/2}$ as $n \to \infty$, which further implies $\|D_{S_0}\|_1$ and $\|D_{S_0^c}\|_1$ are bounded from the quantity as well.

THEOREM 5.1. *Given that Assumptions 1 to 3 hold, if the number of covariates $p \propto n^{\alpha}$, $0 < \alpha < 1/2$, and $\tau_3 \propto n^{-1}$, then we have*

$$\mathbb{P}(E_{0,\tau_3}) \to 1$$

as $n \to \infty$.

The proof is given in Appendix C of Supplementary Material [35]. The proof will start by exploring the KKT conditions associated to the minimization problem stated in (5.1). Note that in Theorem 5.1 we do not assume that the Irrepresentable Condition should hold. Indeed, as stated in Corollary C1 in Appendix C, even if the Irrepresentable Condition is violated, Theorem 5.1 will still hold given that some mild condition is imposed.

5.2. *Posterior model consistency.* We give notation definitions first. Let $y^n = (y_1, y_2, \dots, y_n)$. The notation y^n emphasizes the fact that the number of entries in the observed response vector y is n. Further let $\mathcal{M}_S$ denote the model characterized by the index set S. Under a Bayesian framework, $\mathcal{M}_S$ usually refers to the sampling density, and posterior model consistency is defined as $\mathbb{P}(\mathcal{M}_{S_0}|y^n) \to 1$ as $n \to \infty$, where $\mathcal{M}_{S_0}$ can be seen as the "true model," or the true sampling density that a sample comes from. The posterior model consistency states that the posterior probability will put all its mass on $\mathcal{M}_{S_0}$ as the sample size goes to infinity. Note that in general, multiple true models are allowed under Bayesian frameworks, that is, S_0 may not be unique. However, for simplicity we only pay attention on the situation in which there is only one true model.

Note that the posterior probability $\mathbb{P}(\mathcal{M}_{S'}|y^n)$ with $S' \in \mathcal{S}$ can be expressed in terms of Bayes factors by

$$\begin{aligned}\mathbb{P}(\mathcal{M}_{S'}|y^n) &= \frac{f(y^n|\mathcal{M}_{S'})f(\mathcal{M}_{S'})}{\sum_{S\in\mathcal{S}} f(y^n|\mathcal{M}_S)f(\mathcal{M}_S)} \\ &= \frac{\mathrm{BF}(\mathcal{M}_{S'}, \mathcal{M}_{S_0}; y^n)f(\mathcal{M}_{S'})}{\sum_{S\in\mathcal{S}} \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n)f(\mathcal{M}_S)}.\end{aligned} \tag{5.3}$$

The formulation (5.3) implies that the event $\mathbb{P}(\mathcal{M}_{S_0}|y^n) \to 1$ is equivalent to the events $\mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}) \to 0$ for all $S \in \mathcal{S}$ and $S \neq S_0$, given that the probability $f(\mathcal{M}_{S_0})$ is bounded from zero. It turns out that to examine whether the posterior probability is consistent at the true model $\mathcal{M}_{S_0}$ is the same as to examine whether Bayes factors between other models and the true model will approach to zero or not.

Here, we make some assumptions on the Bayesian formulation (3.1) before stating the main result of posterior model consistency.

ASSUMPTION 4. The prior probability on the true model $\mathcal{M}_{S_0}$ is bounded away from zero, that is, $f(\mathcal{M}_{S_0}) > 0$.

ASSUMPTION 5. We assume $\lambda^{-1}\tau_2 < \infty$.

ASSUMPTION 6. The condition

$$\frac{(y^n)^T y^n}{\Lambda_{\min}(X_{S_0}X_{S_0}^T)} < (y^n)^T(X_S X_S^T + \lambda I)^{-1} y^n \tag{5.4}$$

holds for all $n \in \mathbb{N}^+$, $S \in \mathcal{S} \setminus S_0$ and $0 \leq \lambda < \infty$.

Assumption 4 states that the true model should always have positive mass under the prior. It is a reasonable assumption since otherwise by Bayes' theorem the posterior probability of $\mathcal{M}_{S_0}$ will be zero. In addition, Assumption 6 is a technical condition which ensures that the ratio $(y^n)^T(X_{S_0}X_{S_0}^T + \lambda I)^{-1}y^n/[(y^n)^T(X_S X_S^T +$

$\lambda I)^{-1}y^n]$ is smaller than 1. The assumption will be useful in proving the convergence of the Bayes factor $\mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n)$.

THEOREM 5.2. *Given Assumptions 4 to 6 hold and the number of covariates* $p \propto n^\alpha$, *the inequality*

$$\mathbb{P}(\mathcal{M}_{S_0}|y^n) \geq 1 - c_3 \exp\left\{-\frac{n^\alpha}{2}(n^{1-\alpha}\xi - c_{11}\log 4)\right\}$$

will hold for some constants $0 \leq c_3 < \infty$, $\xi > 0$, $0 < c_{11} < \infty$ *and* $n^* > 0$ *for all* $n > n^*$. *Therefore, for* $0 < \alpha < 1$, $\mathbb{P}(\mathcal{M}_{S_0}|y^n) \to 1$ *as* $n \to \infty$.

The proof of Theorem 5.2 is given in Appendix D of Supplementary Material [35].

5.3. *Estimation consistency.* Using (5.1), the BAVA-MIO estimator can be defined as $\widehat{\beta}_{\mathrm{BMIO}} = \lim_{\tau_3 \to 0} \widehat{\beta}^{\tau_3}$. Now we deal with the estimation consistency of $\widehat{\beta}_{\mathrm{BMIO}}$ under a frequentist's framework. We will derive an asymptotic bound for the l_2 distance between $\widehat{\beta}_{\mathrm{BMIO}}$ and β_0 and show that the l_2 distance converges to 0 as $n \to \infty$. Let β_0 denote the coefficient vector corresponding to the true model $\mathcal{M}_{S_0}$. Define the expected l_2 distance between estimator $\widehat{\beta}$ and the true coefficient β_0 by

$$\mathbb{E}_Y[\|\widehat{\beta} - \beta_0\|_2^2] = \int \|\widehat{\beta} - \beta_0\|_2^2 f(y|\beta_0)\, dy,$$

where $f(y|\beta_0)$ is the sampling density parametrized by the true parameter β_0. Before deriving the asymptotic result, we make some finite moment assumptions on the true parameters (β_0, σ^2).

ASSUMPTION 7. There exist finite constants $c_4 > 0$ and $c_5 > 0$ such that $\beta_{0,j}^2 < c_4$ for $j = 1, 2, \ldots, p$, and $\sigma^2 < c_5$.

Assumption 7 ensures that parameter β_0 and parameter σ^2 are bounded away from above as the sample size n goes large.

THEOREM 5.3. *Given Assumptions 1, 3 and 7 hold and the number of covariates* $p \propto n^\alpha$ *with* $\alpha > 0$, *the inequality*

$$\mathbb{P}(\|\widehat{\beta}_{\mathrm{BMIO}} - \beta_0\|_2^2 > \xi_n) \leq c_{13} \exp\{-\log(n^{1-\alpha}\xi_n)\} \tag{5.5}$$

will hold for some positive finite constant c_{13} *and* ξ_n. *Assume* $\xi_n \geq 0$ *is decreasing with* n, *that is,* $\xi_n \to 0$ *as* $n \to \infty$. *Let* $\xi_n \propto n^{-\alpha^*}$ *for some* $\alpha^* > 0$. *Then with the condition* $0 < \alpha^* < \alpha < 1/2$, $\mathbb{P}(\|\widehat{\beta}_{\mathrm{BMIO}} - \beta_0\|_2^2 > \xi_n) \to 0$ *as* $n \to \infty$.

The proof of Theorem 5.3 is given in Appendix E of Supplementary Material [35].

6. An extension to generalized linear models.

Here we extend the proposed method, the BAVA-MIO, to parameter estimation in the generalized linear models. Consider the density of the exponential family

$$f(y|\theta,\varphi)=\exp\left\{\frac{y\theta-b(\theta)}{\varphi}+d(y,\varphi)\right\}, \tag{6.1}$$

where θ is a parameter characterizing mean of the distribution and φ is a parameter characterizing dispersion of the distribution. Under the exponential family (6.1), variable Y has properties such that $\mathbb{E}(Y|\theta,\varphi)=b'(\theta)$, $\operatorname{Var}(Y|\theta,\varphi)=b''(\theta)\varphi$. Now let $\nu=b'(\theta)$. For a generalized linear model, there exists a link function η such that $\eta(\nu)=x^T\beta$. The link function gives a flexible connection between the mean ν and the predictor $x^T\beta$, and a valid regression can be formulated under this parametrization. In addition, ν is parametrized by θ, therefore by inverse mapping, we can express θ as a function of $x^T\beta$. We write $\theta=\theta(x^T\beta)$.

For a practical inference concern, we will not assign a prior on φ in the following Bayesian hierarchical formulation. We only assign priors on regression coefficients β and covariate indices γ. The inference concern arises from the fact that the estimation of φ is dependent on the function $d(y,\varphi)$, and in general, $d(y,\varphi)$ is case dependent. We will launch an investigation on how to assign a prior on φ in the future, but at present we only focus on inference based on priors on β and γ. Now consider the logarithm of the joint density function

$$\begin{aligned}-\log f(\beta,\gamma|X,y,\varphi,\lambda,\kappa)={}&-\sum_{i=1}^{n}\left\{\frac{y_i\theta_i(x_i^T\beta)-b[\theta_i(x_i^T\beta)]}{\varphi}+d(y_i,\varphi)\right\}\\&+\frac{\lambda}{2\varphi}\sum_{j=1}^{p}\gamma_j\beta_j^2\\&+\frac{1}{2}\sum_{j=1}^{p}\gamma_j\log\left\{\frac{2\pi\varphi(1-\kappa)^2}{\lambda\kappa^2}\right\}+\text{const.}\end{aligned} \tag{6.2}$$

The first term in (6.2) is the logarithm of joint sampling density over $i=1,2,\dots,n$, and the second and third terms are logarithms of the priors on β and covariate indices γ, respectively. To modify (6.2) for BAVA-MIO estimation, we first multiply (6.2) with φ. We then apply a majorization–minimization technique to obtain an approximation to the l_0 norm penalty. The BAVA-MIO estimator of β is defined as the minimizer of the approximate objective, which can be obtained by the following iteration scheme:

$$\begin{aligned}\widehat{\beta}^{(m+1)}=\arg\min\Bigg\{&-\sum_{i=1}^{n}\{y_i\theta_i(x_i^T\beta)-b[\theta_i(x_i^T\beta)]\}\\&+\frac{\lambda}{2}\|\beta\|_2^2+\rho\|\widehat{\phi}^{(m)}\beta\|_1\Bigg\},\end{aligned} \tag{6.3}$$

where $\rho = \varphi[\log 2\pi\varphi(1-\kappa)^2(\lambda\kappa^2)^{-1}]/2$ and $\widehat{\phi}^{(m)} = \lim_{\tau_3\to 0}[\log(1+\tau_3^{-1}) \times (|\widehat{\beta}^{(m)}| + \tau_3)]^{-1}$. Further, by differentiating (6.3) with respect to β, and setting the derivatives to zero, we obtain the subgradient equations of β, which are given by

$$-X^T W r + \lambda\beta + g_\beta \rho \widehat{\phi}^{(m)} = 0, \tag{6.4}$$

where $r = (y-\nu)\eta'(\nu)$, $W = \operatorname{diag}\{[\eta'(\nu_1)^2]b''(\theta_1), \dots, [\eta'(\nu_n)^2]b''(\theta_n)\}^{-1}$, $\nu = (\nu_1, \nu_2, \dots, \nu_n)$ with $\nu_i = b'(\theta_i)$ and $g_\beta = (g_{\beta_1}, g_{\beta_2}, \dots, g_{\beta_p})$ is the subgradient vector of $\|\beta\|_1$ such that $g_{\beta_j} = 1$ if $\beta_j > 0$, $g_{\beta_j} = -1$ if $\beta_j < 0$ and $g_{\beta_j} \in [-1, 1]$ if $\beta_j = 0$. The term $X^T W r$ in (6.4) is a standard result in parameter estimation of the generalized linear models, and its derivation can be found in [26]. The term $X^T W r$ allows us to formulate an iteration scheme to approximate the solution of the subgradient equations (6.4). Here we will use the iteration scheme

$$(\widehat{\beta}^*)^{(m+1)} = \arg\min_\beta \left\{ \frac{1}{2}\|U^{(m)}(z^{(m)} - X\beta)\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2 + \rho\|\widehat{\phi}^{(m)}\beta\|_1 \right\}, \tag{6.5}$$

where

$$z^{(m)} = r^{(m)} + \eta^{(m)},$$
$$U^{(m)} = (W^{1/2})^{(m)},$$

to approximates the solution of the subgradient equations (6.4). The jth element of the iteration scheme (6.5) can be obtained by further carrying out the following soft-thresholding scheme coordinatewise:

$$(\tilde{\beta}_j^*)^{(m+1,l+1)} = \left(\sum_{i=1}^n w_{ii}^{(m)} x_{ij}^2 + \lambda\right)^{-1} \mathrm{ST}\left(\sum_{i=1}^n x_{ij} w_{ii}^{(m)} \tilde{v}_{i,-j}^{(m,l)}, \rho\tilde{\phi}_j^{(m)}\right), \tag{6.6}$$

where $w_{ii}^{(m)}$ is the ith diagonal term of $W^{(m)}$, $\tilde{v}_{i,-j}^{(m,l)} = z_i^{(m)} - \sum_{j'\neq j} x_{ij}\tilde{\beta}_{j'}^*$ with $\tilde{\beta}_{j'}^* = (\tilde{\beta}_{j'}^*)^{(m+1,l+1)}$ for $j' = 1, 2, \dots, j-1$ and $\tilde{\beta}_{j'}^* = (\tilde{\beta}_{j'}^*)^{(m,l)}$ for $j' = j+1, j+2, \dots, p$, and $\mathrm{ST}(a, b)$ is the soft-thresholding operator defined by $\mathrm{ST}(a, b) = \operatorname{sign}(a)(|a| - b)_+$.

We now conduct a simulation study to assess the performance of the BAVA-MIO estimation. We take logistic regression as the example. For the true model, we assume $Y_i \sim \mathrm{Bernoulli}(\zeta_i)$, where ζ_i is parametrized in terms of predictor $x_i^T\beta$ via the link function $\log[(\zeta_i)/(1-\zeta_i)]$. We further let the number of covariates $p = 120$. For the regression coefficients β, we generate the jth entry β_j from Normal(0, 1) for $j = 1, 2, \dots, 10$, and let the rest of 110 $\beta_j' s$ equal to zero. We simulate covariate vector x_i i.i.d. from $\mathrm{MVN}(0, \Sigma_X)$. We consider three Σ_X's, the same as those described in Section 4.1, to generate the covariate vectors. With β and x_i, we simulate Y_i from Bernoulli(ζ_i) for the cases of $n = 100$ and $n = 200$. With three different values for Σ_X and two different values for n, we have six scenarios in the simulation study. For each scenario, we generate 100 simulation runs.

TABLE 2

Results of BAVA-MIO GLM estimation. Each value is calculated by averaging over 100 *simulation runs, and the corresponding standard error is given in the bracket. BMIO-CV: the BAVA-MIO estimation using ten-fold cross validation; lasso: the lasso estimation. The top panel: covariance matrix with off-diagonal terms equal to* 0; *The middle panel: covariance matrix with off-diagonal terms equal to* 0.5; *The bottom panel: covariance matrix with off-diagonal terms following a specified covariance structure*

	n	PMSE	l_2-dis	S-FPR	$\lvert\widehat{S}\rvert$
BMIO-CV	100	0.186 (0.004)	0.039 (0.002)	0.305 (0.031)	11.39 (1.788)
GLM-lasso	100	0.173 (0.003)	0.044 (0.004)	0.680 (0.012)	21.34 (1.050)
BMIO-CV	200	0.145 (0.003)	0.018 (0.001)	0.176 (0.021)	7.54 (0.549)
GLM-lasso	200	0.144 (0.002)	0.026 (0.001)	0.694 (0.011)	27.65 (1.141)
BMIO-CV	100	0.180 (0.004)	0.055 (0.003)	0.455 (0.031)	13.86 (1.841)
GLM-lasso	100	0.169 (0.003)	0.052 (0.003)	0.654 (0.018)	17.54 (0.966)
BMIO-CV	200	0.157 (0.003)	0.027 (0.002)	0.327 (0.025)	8.58 (0.564)
GLM-lasso	200	0.154 (0.003)	0.030 (0.001)	0.654 (0.012)	20.67 (0.908)
BMIO-CV	100	0.180 (0.004)	0.046 (0.003)	0.271 (0.028)	8.28 (1.187)
GLM-lasso	100	0.170 (0.003)	0.047 (0.003)	0.668 (0.016)	19.18 (0.914)
BMIO-CV	200	0.152 (0.004)	0.022 (0.001)	0.203 (0.023)	7.19 (0.478)
GLM-lasso	200	0.150 (0.003)	0.027 (0.001)	0.675 (0.014)	23.50 (1.045)

In each simulation run, we apply the BAVA-MIO estimation to fit a logistic regression model. We let hyperparameter $\lambda = n^{-1/2}$ for all estimations. Note that for a Bernoulli variable, a closed form representation for the Bayes factor does not exist, therefore we only use ten fold cross validation for tuning parameter selection. For comparison purposes, we also carry out the lasso estimation using R package “glmnet” and use ten fold cross validation for tuning parameter selection. We collect four performance measures, the same as those described in Section 4.1, at each simulation run. Average values of the four performance measures over the 100 simulation runs are given in Table 2. From these tables we can see that the BAVA-MIO estimation in general has slightly larger values in PMSE than the lasso estimation has, but it gives far fewer number of selected covariates, more accurate results in covariate selection, and in some circumstances, better parameter estimation than the lasso estimation.

7. Real data examples.

In this section, we present two real data analyses. We will apply methods developed in Section 3 and Section 6 to estimate parameters in regression models.

7.1. *Diabetes data.*

The Diabetes data contains a measure on disease progression and 10 covariates: age, sex, the BMI index, blood pressure and six related variables for 442 diabetes patients. In our analysis, each covariate has

TABLE 3

Estimation results based on the Diabetes data. The value in the bracket is the inclusion probability of the covariate based on the 100 *subsampling estimations. For g-prior, hyper-g and BIC, the value in the bracket is the posterior inclusion probability of the covariate*

Name	**BMIO-BF**	**BMIO-CV**	**g-prior**	**hyper-g**	**BIC**
age	0.00 (0.01)	0.00 (0.30)	0.00 (0.11)	0.00 (0.33)	0.00 (0.05)
sex	−11.23 (0.49)	−11.08 (0.64)	−10.64 (0.99)	−8.02 (0.97)	−10.71 (0.98)
bmi	24.92 (1.00)	25.06 (1.00)	24.96 (1.00)	19.00 (1.00)	25.37 (1.00)
map	15.54 (0.86)	15.01 (0.92)	15.29 (1.00)	11.55 (1.00)	15.53 (1.00)
tc	0.00 (0.03)	0.00 (0.33)	−16.62 (0.71)	−13.54 (0.75)	0.00 (0.57)
ldl	0.00 (0.10)	0.00 (0.28)	8.51 (0.50)	6.51 (0.59)	0.00 (0.38)
hdl	−13.76 (0.69)	−11.20 (0.81)	0.00 (0.49)	0.00 (0.57)	−7.29 (0.57)
tch	0.00 (0.01)	0.00 (0.27)	0.00 (0.30)	0.00 (0.48)	0.00 (0.20)
ltg	22.59 (1.00)	25.72 (1.00)	29.13 (1.00)	22.29 (1.00)	28.31 (1.00)
glu	0.00 (0.10)	3.44 (0.47)	0.00 (0.17)	0.00 (0.41)	0.00 (0.07)

been rescaled to have mean zero and variance 1, and the response variable has been centered around its mean. All estimations are based on the rescaled covariates and centered response variable. For hyperparameters, we let $(\tau_1, \tau_2) = (1, 1)$ and $\lambda = 0.2 \times \sqrt{p \log(p)/n} \approx 0.049$. We perform two BAVA-MIO estimations. The first one uses the Bayes factor (BMIO-BF) while the second one uses ten fold cross validation (BMIO-CV) for tuning parameter selection. The results are shown in the first two columns of Table 3. From the results, we can see the BMIO-BF estimation leads to a covariate selection sparser than its counterpart using ten fold cross validation. We also run another 100 estimations based on sampling half of the 442 subjects without replacement to calculate the inclusion probabilities for the 10 covariates. For each covariate, the inclusion probability is defined as the proportion of occurrences of nonzero estimated values appearing in the 100 subsampling estimations. We compare the results from the BAVA-MIO estimations with the results from three other estimation approaches: g-prior, hyper-g and BIC. All the three estimations are carried out using R package "BAS," which is developed by Clyde, Ghosh and Littman [5] as the companion software to the paper of Liang et al. [24]. These results are shown in the last three columns of Table 3. For the three estimations using the BAS package, we report the models estimated with the highest marginalized likelihood. The results show that the estimation based on BAVA-MIO using the Bayes factor has relative sparse covariate selection among the five proposed approaches. Among the 10 inclusion probabilities estimated via the BMIO-BF estimation, only four are above 0.5, compared to five for the BMIO-CV estimation, six for the g-prior and the BIC estimations, and seven for the hyper-g estimation.

TABLE 4
Classification results for Golub's gene expression data

Method	**CV-error**	**Test-error**	**# of genes**
Golub et al. [15]	$3/38$	$4/34$	50
Elastic Net (Zou and Hastie [41])	$3/38$	$0/34$	45
l_1-pen GLM (Park and Hastie [31])	$1/38$	$2/34$	23
SIS-SCAD-LD (Fan and Lv [8])	$0/38$	$1/34$	16
FAIR (Fan and Fan [6])	$1/38$	$1/34$	11
BMIO-CV I	$1/38$	$1/34$	8
BMIO-CV II	$1/38$	$1/34$	9
BMIO-CV III	$1/38$	$0/34$	23

7.2. *Golub's Leukemia data.* The Leukemia gene expression data, adopted from R package "golubEsets," is originally from [15]. It consists of gene expression profiles for 72 Leukemia patients, of which 47 are diagnosed with acute lymphoblastic leukemia (ALL) and 25 are diagnosed with acute myeloid leukemia (AML). Each profile has 7,129 gene expression values measured by Affymetrix Hgu6800 chips. The data set is further divided into the training set, which consists of 27 ALL patients and 11 AML patients, and the test set, which consists of 20 ALL patients and 14 AML patients. Our aim is to identify a patient's disease type with a small set of genes. The data set is processed as follows. The disease type is labeled with 0 for the acute lymphoblastic leukemia and 1 for the acute myeloid leukemia. Each covariate is first rescaled to have a range greater than or equal to zero. Then it is under a suitable logarithm transform before rescaled again to have mean 0 and variance 1. For the classification rule construction, we apply the BAVA-MIO estimation to fit logistic regression models with the training data. We parametrize hyperparameter $\lambda = \lambda^* \sqrt{p \log p/n}$ and perform three estimations with $\lambda^* = 0.05$, 0.1 and 0.5. The tuning parameter is selected via five fold cross validation and the resulting estimates are termed BMIO-CV I, BMIO-CV II and BMIO-CV III, respectively. With estimated regression coefficients, we calculate the label probability for each patient, and classifying those with label probabilities smaller than 0.5 to the acute lymphoblastic leukemia group, and those with label probabilities greater than 0.5 to the acute myeloid leukemia group. The corresponding classification results are reported in Table 4, along with classification results on the same data set done by Golub et al. [15] and four other estimation approaches [6, 8, 31, 41] aiming to tackle high-dimensional classification problems. The results show that BAVA-MIO-based classification rules tend to use less numbers of genes in identifying a patient's disease type. However, even with smaller numbers of genes, the BAVA-MIO-based classification rules can still generate results that are comparable with those provided by other benchmark methods.

8. Concluding remarks. One important issue to which we did not pay much attention is the impacts of hyperparameters on estimation results. Here we provide some possible modifications in addressing this issue. First, an equally spaced grid may be constructed for hyperparameter λ so that the estimation procedure can be carried out along the grids on λ and $\Psi(\kappa)$. Another possible modification is to drop the prior assumption on σ^2 and treat it as a constant. In this approach the impact of σ^2 on parameter estimation can be dealed together with the tuning parameter $\Psi(\kappa)$. This approach has been adopted in Section 6 for parameter estimation in the generalized linear models.

Acknowledgments. We thank the Associate Editor and the reviewer for their invaluable comments. We are grateful to Dr. Chun-houh Chen for his encouragement and helpful suggestions. We also thank Professor Yuan-chin Chang, Dr. Ting-Li Chen, Dr. Kai-Ming Chang and Mr. Yu-Min Yen for helpful comments.

SUPPLEMENTARY MATERIAL

Supplement File (DOI: 10.1214/11-AOS884SUPP; .pdf). In Supplementary Material, we provide brief discussions on the log-sum function, connections with other approaches, derivation of the soft-thresolding operator, and proofs of Theorems 5.1, 5.2 and 5.3.

REFERENCES

[1] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[2] CANDÉS, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[3] CANDÉS, E. J., WAKIN, M. B. and BOYD, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14** 877–905. MR2461611

[4] CLYDE, M. and GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 681–698. MR1796285

[5] CLYDE, M., GHOSH, J. and LITTMAN, M. (2011). Bayesian adaptive sampling for variable selection and model averaging. *J. Comput. Graph. Statist.* **20** 80–101.

[6] FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

[7] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[8] FAN, J. and LV, J. (2008). Sure independence screeing for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[9] FRIEDMAN, J., HASTIE, T., HÖLFING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

[10] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33** 1–22.

[11] GENKIN, A., LEWIS, D. D. and MADIGAN, D. (2007). Large scale Bayesian logistic regression for text categorization. *Technometrics* **49** 291–304. MR2408634

[12] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

[13] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

[14] GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

[15] GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J., COLLER, H., LOH, M., DOWNING, J., CALIGIURI, M., BLOOMFIELD, C. and LANDER, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

[16] GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188.

[17] HANS, C. (2009). Bayesian lasso regression. *Biometrika* **96** 835–845. MR2564494

[18] HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

[19] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100** 764–780. MR2201009

[20] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

[21] JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560

[22] KNIGHT, K. and FU, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[23] LI, Q. and LIN, N. (2010). The Bayesian elastic net. *Bayesian Anal.* **5** 151–170.

[24] LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2007). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

[25] MAZUMDER, R., FRIEDMAN, J. and HASTIE, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* To appear.

[26] MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York. MR0727836

[27] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

[28] MEINSHAUSEN, N. (2007). Relaxed lasso. *Comput. Statist. Data Anal.* **52** 374–393. MR2409990

[29] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[30] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1032. MR0997578

[31] PARK, M. Y. and HASTIE, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 659–677. MR2370074

[32] PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

[33] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

[34] WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601

[35] YEN, T.-J. (2011). Supplement to "A majorization–minimization approach to variable selection using spike and slab priors." DOI:10.1214/11-AOS884SUPP.

[36] YUAN, M. and LIN., Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[37] YUAN, M. and LIN, Y. (2007). On the nonnegative garrotte estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 143–161. MR2325269

[38] ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[39] ZHAO, P. and YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7** 2541–2564. MR2274449

[40] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[41] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

[42] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

[43] ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. MR2533470

INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
128 ACADEMIA ROAD, SECTION 2
TAIPEI 115
TAIWAN
E-MAIL: tjyen@stat.sinica.edu.tw