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

BY TSO-JUNG YEN∗

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:

\[ Y_i = x_{i1} \beta_1 + x_{i2} \beta_2 + \cdots + x_{ip} \beta_p + \epsilon_i, \]

where $Y_i$ is the response variable for the $i$th subject, $x_{ij}$ is the $j$th covariate for the $i$th subject, $\beta_j$ is the corresponding regression coefficient, and $\epsilon_i$ 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 $\hat{S}$ of indices $\{1, 2, \cdots, p\}$ so that for $j \in \hat{S}$, the covariates $x_{ij}$’s and estimated coefficients $\hat{\beta}_j$’s are scientifically meaningful in interpretation, and estimates $\hat{y}_i = \sum_{j \in \hat{S}} x_{ij} \hat{\beta}_j$ 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

∗Postdoctoral Fellow, Institute of Statistical Science, Academia Sinica.
AMS 2000 subject classifications: Primary 62H12; secondary 62F15, 62J05.
Keywords and phrases: MAP estimation, $l_0$ norm, majorization-minimization algorithms, Irrepresentable Condition.
previously been studied by Genkin et al. [11] on logistic regression models with Laplace priors. 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, 31, 17, 23, 16] and empirical Bayes methods [12, 4, 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 [20, 19], 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, i.e. 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 non-zero 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 \([7, 41, 40, 42, 9, 39, 35, 27, 2, 28, 37]\). 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 [39], 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 [38] 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 [39] and Yuan and Lin [36]. 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 $i$th row of $X$ and $x_{ij}$ denote the $ij$th entry of $X$. The transpose of $X$ is denoted by $X^T$. Let $y = (y_1, y_2, \cdots, y_n)$ denote the realization of random vector $Y = (Y_1, Y_2, \cdots, Y_n)$ and $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ denote the regression coefficient vector. Let $I_{p \times p}$ denote the $p \times p$ identity matrix. For a $p$-dimensional vector $a = (a_1, a_2, \cdots, a_p)$, 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 $l_\infty$ norm by $||a||_\infty = \max_j |a_j|$, and the $l_0$ norm by $||a||_{l_0} = \sum_{j=1}^{p} I(a_j \neq 0)$, where $I(a_j \neq 0)$ is an index variable such that $I(a_j \neq 0) = 1$ if $a_j \neq 0$, and $I(a_j \neq 0) = 0$ otherwise. The probability density of a random variable $Z$ conditional on $\theta$ is denoted by $f(z | \theta)$. We define $S = \{j : \beta_j \neq 0, j = 1, 2, \cdots, p\}$, i.e. the index set of non-zero valued coefficients in $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$. 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 non-zero value. Based on this observation, we assign an index variable $\gamma_j$ to each covariate and define that $\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ if $\beta_j = 0$. In practice we may
write \( \gamma_j = \mathbb{I}(\beta_j \neq 0) \). With the definition of \( \gamma_j \), the regression model (1.1) has an equivalent representation given by

\[
Y_i = \sum_{j=1}^{p} x_{ij} \gamma_j \beta_j + \epsilon_i.
\]

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

### 3.1. The Bayesian formulation

Under a Bayesian framework, we assume

\[
Y_i | x_i, \beta, \gamma, \sigma^2 \sim \text{Normal} \left( \sum_{j=1}^{p} x_{ij} \gamma_j \beta_j, \sigma^2 \right) \quad \text{for } i = 1, 2, \cdots, n,
\]

\[
\beta_j | \sigma^2, \gamma_j, \lambda \sim \gamma_j \text{Normal}(0, \sigma^2 \lambda^{-1}) + (1 - \gamma_j) \mathbb{I}(\beta_j = 0)
\]

\[
\sigma^2 | \tau_1, \tau_2 \sim \text{Inverse-Gamma}(\tau_1, \tau_2),
\]

\[
\gamma_j | \kappa \sim \text{Bernoulli}(\kappa) \quad \text{for } j = 1, 2, \cdots, p.
\]

(3.1)

The prior distribution of \( \beta_j \) given in (3.1) is the spike and slab originally proposed by Mitchell and Beauchamp [30]. It implies that conditional on \( \gamma_j = 0 \), \( \beta_j \) is equal to 0 with probability one, and conditional on \( \gamma_j = 1 \), \( \beta_j \) follows a normal distribution with mean 0 and variance \( \sigma^2 \lambda^{-1} \). The Bernoulli prior on \( \gamma_j \) says that if only prior information is available, \( \gamma_j \) will have probability \( \kappa \) 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 \( \beta_j \) 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 \( \beta, \gamma \) and \( \sigma^2 \) can be expressed as

\[
f(\beta, \gamma, \sigma^2 | X, y, \lambda, \tau_1, \tau_2, \kappa)
\]

\[
\propto f(y | X, \sigma^2, \beta, \gamma) f(\beta | \sigma^2, \gamma, \lambda) f(\sigma^2 | \tau_1, \tau_2) f(\gamma | \kappa).
\]

(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

\[
(\hat{\beta}, \hat{\gamma}, \hat{\sigma}^2) = \arg \min_{\beta, \gamma, \sigma^2} \{-2 \log f(\beta, \gamma, \sigma^2 | X, y, \lambda, \tau_1, \tau_2, \kappa)\},
\]

(3.3)
i.e. 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

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

(3.4)

Here we have used an equivalent representation Normal(0, \sigma^2 \lambda^{-1}) \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 \( \sigma^2 \) and multiply (3.4) with \( \sigma^2 \) in the following discussion. Given that \( \sigma^2 \) is fixed, the function (3.4) has some meaningful interpretations in terms of regularization estimation on \( \beta \). 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 \( \gamma \). 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 \( \kappa \) 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 \( \beta \), 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 \( \sigma^2 \) can be derived by first making a derivative of (3.4) with respect to \( \sigma^2 \), setting the derivative to zero, and then solving the equation for \( \sigma^2 \). The estimation of \( \beta \) is further carried out given \( \sigma^2 \) is fixed. With fixed \( \sigma^2 \) and the interpretations of regularization estimation given above,
(3.4) has an equivalent representation given by

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

Note that here we have multiplied (3.4) with \( \sigma^2 \). 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

\[
\hat{\sigma}^{(m+1)} = \frac{\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\hat{\gamma}^{(m)}_j \hat{\beta}^{(m)}_j)^2 + \lambda \sum_{j=1}^p \hat{\gamma}^{(m)}_j (\hat{\beta}^{(m)}_j)^2 + 2\tau_2}{n + \sum_{j=1}^p \hat{\gamma}^{(m)}_j + 2\tau_1 + 2},
\]

\[
\hat{\beta}^{(m+1)} = \arg \min_{\beta} L(\beta; \lambda, \rho_{\lambda,\kappa,\hat{\sigma}^{(m+1)}}),
\]

\[
\hat{\gamma}^{(m+1)} = (I(\hat{\beta}^{(m+1)}_1 \neq 0), I(\hat{\beta}^{(m+1)}_2 \neq 0), \ldots, I(\hat{\beta}^{(m+1)}_p \neq 0)).
\]

(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 begin the approach by noting that, mathematically the \( l_0 \) norm on a \( p \)-dimensional vector \( \beta \) can be expressed as

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

(3.7)

which can be verified by seeing (3.7) as a function of \( \tau_3 \) and using l’Hopital’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. With representation (3.7), the objective function (3.5) can be re-expressed as

\[
L(\beta; \lambda, \rho_{\lambda,\kappa,\sigma^2}) = ||y - X\beta||^2_2 + \lambda ||\beta||^2_2
\]

\[
+ \rho_{\lambda,\kappa,\sigma^2} \left\{ \lim_{\tau_3 \to 0} \sum_{j=1}^p \frac{\log(1 + \tau_3^{-1} |\beta_j|)}{\log(1 + \tau_3^{-1})} \right\} + \text{const}. \]

(3.8)

If \( \tau_3 \) 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 \( \tau_3 \) 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 \( \beta \), the combinatorial nature of \( ||\beta||_0 \) is relaxed. However, the term \( \log(1 + \tau_3^{-1} |\beta_j|) \) is not convex in \( \beta_j \), and
replacing $||\beta||_0$ with (3.7) in (3.5) still makes objective function (3.8) remain non-convex. 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:

$$
\begin{align*}
    h(\theta; \theta^{(l)}) &\geq g(\theta) \quad \text{for all } \theta, \\
    h(\theta^{(l)}; \theta^{(l)}) &= g(\theta^{(l)}).
\end{align*}
$$

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

$$
\begin{align*}
    g(\theta^{(l)}) = h(\theta^{(l)}; \theta^{(l)}) \geq h(\theta^{(l+1)}; \theta^{(l)}) \geq g(\theta^{(l+1)}),
\end{align*}
$$

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 $\theta$ for $\theta > 0$, therefore the inequality

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

holds for all $\theta > 0$ and $\theta' > 0$. Note that the left hand side of (3.11) is convex in $\theta$. 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:

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

where

$$
h_2(\beta; \beta') = \lim_{\tau_3 \to 0} \rho_3 \sum_{j=1}^{p} \left( \log \left( 1 + \tau_3^{-1} |\beta_j'| \right) + \frac{|\beta_j| + \tau_3}{|\beta_j'| + \tau_3} - 1 \right).
$$
Proof of Proposition 3.1. Let \( h_1(\beta) = ||y - X\beta||_2^2 + \lambda||\beta||_2^2 \). Assume \( \beta^{(l+1)} \) minimizes \( L''(\beta; \lambda, \rho, \kappa, \sigma^2) \) given \( \beta' = \beta^{(l)} \). Then with (3.7) and the inequality (3.11), the quantity \( L'(|\beta^{(l+1)}|; \lambda, \rho, \kappa, \sigma^2) \) can be bounded in a way such that
\[
L'(|\beta^{(l+1)}|; \lambda, \rho, \kappa, \sigma^2) = h_1(|\beta^{(l+1)}|) + \rho \log(1 + \tau^{-1} ||\beta^{(l+1)}||)
\]
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 \( L(\beta; \lambda, \rho, \kappa, \sigma^2) \), 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), \( \hat{\beta}^{(m+1)} \) can be obtained by carrying out the following iteration scheme:
\[
\hat{\beta}^{(m+1,l+1)} = \arg\min_\beta \left\{ ||y - X\beta||_2^2 + \lambda||\beta||_2^2 + \rho \log(1 + \tau^{-1} ||\hat{\beta}^{(m+1,l)}||) \right\}
\]
over index \( l \), where \( \hat{\phi}^{(m+1,l)} = \lim_{\tau_3 \to 0} \log(1 + \tau^{-1} ||\hat{\phi}^{(m+1,l)}|| + \tau_3)^{-1} \). The procedure of using iteration scheme (3.15) in obtaining the minimizer for the objective function (3.5), or equivalent (3.8), is called the BAVA-MIO (BAyesian VAriable selection using a Majorization-mInimiz ation 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 \( \hat{\phi}^{(m+1,l)} \), a convex function of \( \beta_j \), therefore given \( ||y - X\beta||_2^2 + \lambda||\beta||_2^2 \) is convex in \( \beta \), the whole objective function in (3.15) will be convex in \( \beta \), 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. [10]. 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 $j$th coordinate is given by

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

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

### 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. For example, in the marginalized likelihood, we wish to see $\lambda \tau_2 \to 0$ as $n \to \infty$. In this sense, we may let $\lambda \tau_2 \propto (p \log p)/n$. In addition, we let $\tau_1 = \tau_2 + 1$, so that the prior expectation of $\sigma^2$ 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 $\kappa$. In some circumstances data-driven empirical Bayes approaches such as those developed in [12] are proposed to obtain $\kappa$, while in other circumstances full Bayesian methods that assign priors on $\kappa$ are proposed. For example, please see [24]. Unlike...
previously proposed approaches, in which single point estimates were obtained for \( \kappa \), we adopt an approach by specifying a feasible region for the function

\[
\psi(\kappa) = \frac{1}{2}\log(2\pi\sigma^2\lambda^{-1}(1-\kappa^2/\kappa^2))\hat{\phi}(0),
\]

and carry out parameter estimation under different values of \( \psi(\kappa) \). Here we have assumed \( \hat{\phi}_j^{(0)} = \hat{\phi}(0) \) for \( j = 1, 2, \cdots, p \). Note that by definition the term \( \hat{\phi}_j^{(0)} \) is a function of the initial value \( \hat{\beta}_j^{(0)} \). 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 \( \kappa \) on the feasible region. Given different values of \( \kappa \), 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 at some stationary points. It can occur in a situation in which we need an initial value \( \hat{\phi}_j^{(1)} \) to run the iteration scheme (3.15). By definition, \( \hat{\phi}_j^{(1)} \) is a function of \( \hat{\beta}_j^{(1)} \), which by definition, is a function of \( \psi(\kappa) \). As pointed out by Candés et al. [3] and Mazumder et al. [25], different \( \hat{\phi}_j^{(1)} \) 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 \( \hat{\phi}_j^{(1)} \), therefore eliminating the possibility that the solution is stuck at 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 [11, 10, 27], and is advocated by [34, 9] 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 \( \beta = (\beta_1, \beta_2, \cdots, \beta_{1000}) \), we let \( \beta_{250} = 2, \beta_{500} = -3.2, \beta_{750} = -1.25, \beta_{1000} = 5.44 \), and \( \beta_j = 0 \) for all \( j \)'s \( \in \{1, 2, \cdots, 1000\} \setminus \{250, 500, 750, 1000\} \). We generate each row of \( X \) independently identically from \( \text{MVN}(0, I_{p \times p}) \), and then calculating \( Y = X\beta + \epsilon \) with \( \epsilon \sim \text{MVN}(0, I_{n \times n}) \). 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

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

where the term \( f(y|\gamma, \tau_1, \tau_2, \kappa, \lambda) \) refers to the marginalized likelihood with \( \beta \) and \( \sigma^2 \) 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

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

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 non-zero 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 non-zero, the corresponding estimation path makes a sharp jump to the non-thresholded 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 non-zero coefficients. In addition, as shown in the right panel of Figure 2, values of the non-zero 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 on the performance of the BAVA-MIO estimation.
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 $Y = X\beta + \epsilon$ with data generated from different simulation schemes. Here $Y$ and $\epsilon$ are $n$-dimensional vectors, $X$ is an $n \times p$ matrix, and $\beta$ is a $p$-dimensional vector. We assume each entry in $\epsilon$ is i.i.d. from $\text{Normal}(0, \sigma_{Y}^{2})$, and each row in the design matrix $X$ is i.i.d. from $\text{MVN}(0, \Sigma_{X})$. Throughout the whole simulation study, we let $p = 120$. For regression coefficient vector $\beta = (\beta_{1}, \beta_{2}, \ldots, \beta_{p})$, we generate $\beta_{j}$ from $\text{Normal}(0, 1)$ for $j = 1, 2, \ldots, 10$ and let $\beta_{j} = 0$ for $j = 11, 12, \ldots, 120$. That is, we have 10 non-zero 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 $\epsilon$. We apply three different $\Sigma_{X}$ 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

![Fig 2. Estimation results for the toy example.](image-url)
signal-to-noise ratio by \( \text{SNR} = \sqrt{\mathbb{E}(\beta^T \Sigma X \beta)/\sigma_Y^2} \). We consider \( \sigma_Y^2 = 10, 1 \) and 0.2 in generating the error term \( \epsilon \). These values correspond to SNR = 1, 3.16 and 7.07, respectively. By using \( X, \beta \) and \( \epsilon \), the response vector \( Y \) is calculated by \( Y = X\beta + \epsilon \). For the number of samples, we consider five values \( n = 40, 80, 120, 160 \) and 200. With three different structures for \( \Sigma X \), three different values for \( \sigma_Y^2 \), 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 \((\lambda, \tau_1, \tau_2) = \left(\frac{1}{\sqrt{n}}, \frac{p \log p}{\sqrt{n}} + 1, \frac{p \log p}{\sqrt{n}}\right)\) for the cases of \( \text{SNR} = 3.16 \) and 7.07. For the case of \( \text{SNR} = 1 \), we use

\[
\begin{align*}
\lambda &= \left(1 - \frac{\text{corr}}{\text{corr}}\right)^2 \frac{p}{n} \log p, \\
\tau_1 &= \left(\frac{\text{corr}}{1 - \text{corr}}\right) \left(\frac{p \log p}{\sqrt{n}}\right)^{\text{corr}/\log n} + 1, \\
\tau_2 &= \frac{1}{\sqrt{n}} \left(\frac{p \log p}{\sqrt{n}}\right)^{1 + \text{corr}/\log n},
\end{align*}
\]

(4.1)

where \( \text{corr} \) 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 \([39]\). 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 \( \lambda \) 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 \( \hat{\beta} \) and the true regression coefficient vector \( \beta \), which is defined by

\[
l_2\text{-dis}(\hat{\beta}) = \sqrt{\frac{\sum_{j=1}^{p} (\hat{\beta}_j - \beta_j)^2}{\sum_{j=1}^{p} \beta_j^2}}.
\]
The second one is the predictive mean squared error of \( \hat{\beta} \) for a test data set, which is defined by

\[
\text{PMSE}(\hat{\beta}) = \frac{\sum_{i=1}^{n_{\text{test}}} (x_{i,\text{test}}^T \hat{\beta} - x_{i,\text{test}}^T \beta)^2}{n_{\text{test}}}.
\]

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

\[
\text{S-FPR} = \frac{\#\{ j \in \hat{S} : \text{sign}(\hat{\beta}_j) \neq \text{sign}(\beta_{\text{true},j}) \}}{|\hat{S}|},
\]

where the sign function \( \text{sign}(.) \) 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 \( n/|S| \), i.e. the ratio between the number of samples and the number of true coefficients with non-zero 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 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 non-zero estimates, as shown in the fourth column of each figure. Moreover, since the BAVA-MIO estimation using the Bayes factor often has fewer numbers of non-zero estimates than other estimation approaches, 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 when 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 some notation definitions. We define \( S_0 = \{ j : \beta_j \neq 0 \} \).
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 non-zero estimates.

0, for some $j \in \{1, 2, \cdots, p\}$ and $S_0^c = \{1, 2, \cdots, p\} \setminus S_0$. Let $\beta_{S_0}$ 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 $\hat{\beta}(n)$ is said to be sign consistent in estimating $\beta$ if the probability of the event \{sign($\hat{\beta}(n)$) = sign($\beta$)\} approaches to 1 as $n \to \infty$. Given the sign consistency holds, the estimated index set $\hat{S}_0 = \{j : \hat{\beta}_j \neq 0\}$ 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, non-zero valued coefficients will have non-zero estimated values, and zero-valued coefficients will be estimated with zero values.

Zhao and Yu [38] showed that if one wants the lasso estimation to achieve the sign consistency, then the design matrices $X$ must satisfy the following
condition:

\[(4.2) \quad ||X_{S_0}^T X_{S_0} (X_{S_0}^T X_{S_0})^{-1} \text{sign}(\beta_{S_0})||_{\infty} < 1,\]

where $\beta_{S_0}$ is the vector of non-zero 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 $X_{S_0}^T X_{S_0} (X_{S_0}^T X_{S_0})^{-1}$ as a least squares estimate for the regression $X_{S_0}$ on $X_{S_0}$. In this sense, the Irrepresentable Condition states that the largest amount of coefficients for the regression $X_{S_0}$ on $X_{S_0}$ should not exceed 1, i.e. $X_{S_0}$ 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 MVN(0, $\Sigma_X$), with $\Sigma_X \sim$
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 non-zero estimates.

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 $\beta$ are generated in a way that the first 5 entries of $\beta$ are i.i.d. from Normal(0, 1), and the rest of 25 entries are set to 0. Note that for some pairs $(X, \beta)$, the Irrepresentable Condition (4.2) will hold, but for some pairs it will not hold. Zhao and Yu defined the Irrepresentable statistic by

$$\text{Irr.stat} = 1 - ||X_{S_0}^T X_{S_0} (X_{S_0}^T X_{S_0})^{-1} \text{sign}(\beta_{S_0})||_\infty$$

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 $(X, \beta)$. In each run we generate $n = 100$ data points. We use $\sigma^2 = 0.05$ to generate the error term $e$, which is corresponding to SNR = 10. We then fit the regression model using the lasso
Fig 6. Scatter plots for the sign probability $\mathbb{P}(\hat{S} = S_0)$ against the Irrepresentable Statistic under different signal-to-noise ratios.

Table 1
The sign probability $\mathbb{P}(\hat{S} = S_0)$ 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 $\tau$ for the correlation calculation.

| Name | SNR = 10 | SNR = 5 | SNR = 2 | SNR = 1 |
|------|---------|---------|---------|---------|
| BMIO | .398 (.030) | .338 (.030) | .077 (.018) | .023 (.006) |
| corr. | .052 | .047 | .180 | .110 |
| Lasso | .314 (.047) | .203 (.030) | .072 (.016) | .022 (.006) |
| corr. | .418 | .232 | .194 | .107 |

estimation and the BAVA-MIO estimation with these data points. For each pair $(X, \beta)$, we calculate the model selection probability $P(\hat{S}_0 = S_0)$ 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 the Irr.stat under signal-to-noise ratios SNR = 10, 5, and 2. From these plots we can 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, i.e. the Irr.stat is smaller than zero. In Section 5 we will provide theoretical results to explain this phenomenon asymptotically. 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 \( S_0, S_0^c, \beta_{S_0}, \beta_{S_0}^c, X_{S_0}, X_{S_0}^c \). Further let \( S \) 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 \( \sigma^2 \) 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

\[
(5.1) \quad \hat{\beta}_{\tau_3} = \arg \min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_2^2 + \rho \sum_{j=1}^{p} \frac{\log(1 + \tau_3^{-1}|\beta_j|)}{\log(1 + \tau_3^{-1})}.
\]

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

\[
(5.2) \quad E_{0, \tau_3} = \left\{ \beta : \text{sign}(\beta_j) = \text{sign}(\hat{\beta}_{\tau_3}^j) \quad \text{for} \quad j = 1, 2, \cdots, p \right\},
\]

i.e. the event of sign consistency for the estimator \( \hat{\beta}_{\tau_3} \) in (5.1). For practical purposes, further define \( C_{S_0} = n^{-1}(X_{S_0}^T X_{S_0} + \lambda I), \quad C_{S_0^c} = n^{-1}X_{S_0}^T X_{S_0^c}, \quad D_{S_0} = n^{-1/2} X_{S_0}^T \epsilon \) and \( D_{S_0^c} = n^{-1/2} X_{S_0^c}^T \epsilon \). 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 \( S \in S \), the maximum eigenvalue \( \Lambda_{\max}(C_{SS}) \) and the minimum eigenvalue \( \Lambda_{\min}(C_{SS}) \) satisfy the following condition:

\[
0 \leq c_1 < \Lambda_{\min}(C_{SS}) \leq \Lambda_{\max}(C_{SS}) \leq c_2 < \infty
\]

**Assumption 2.** For the vector \( X^T \epsilon \), \( ||X^T \epsilon||_1 = O(p) \).
**Assumption 3.** For parameter $\lambda$, we assume $0 \leq \lambda < \infty$. For parameter $\rho$, we assume $0 \leq \rho$ and $\rho n^{-1/2} \rightarrow 0$.

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 $[\Lambda_{\text{max}}(X_S^T X_S) + \lambda] / [\Lambda_{\text{min}}(X_S^T X_S) + \lambda] \leq c_2 / c_1$ is bounded from above for any $S \in S$. Assumption 2 is equivalent to the statement that $n^{-1/2}||X^T \epsilon||_1$ is bounded from a quantity proportional to $\rho n^{-1/2}$ as $n \rightarrow \infty$, 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

$$\mathbb{P}(E_0, \tau_3) \rightarrow 1$$

as $n \rightarrow \infty$.

The proof is given in Appendix C of Supplementary Material. 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 $y^n = (y_1, y_2, \ldots, y_n)$. The notation $y^n$ emphasizes the fact that the number of entries in the observed response vector $y$ is $n$. Further let $M_S$ denote the model characterized by the index set $S$. Under a Bayesian framework, $M_S$ usually refers to the sampling density, and posterior model consistency is defined as $\mathbb{P}(M_{S_0} | y^n) \rightarrow 1$ as $n \rightarrow \infty$, where $M_{S_0}$ 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 $M_{S_0}$ as the sample size goes to infinity. Note that in general, multiple true models are allowed under Bayesian frameworks, i.e. $S_0$ 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 $\mathbb{P}(M_{S'} | y^n)$ with $S' \in S$ can be expressed in terms of Bayes factors by

$$\mathbb{P}(M_{S'} | y^n) = \frac{f(y^n | M_{S'}) f(M_{S'})}{\sum_{S \in S} f(y^n | M_S) f(M_S)} = \frac{\text{BF}(M_{S'}, M_{S_0}; y^n) f(M_{S'})}{\sum_{S \in S} \text{BF}(M_S, M_{S_0}; y^n) f(M_S)}.$$  
(5.3)
The formulation (5.3) implies that the event $P(M_{S_0} | y^n) \rightarrow 1$ is equivalent to the events $BF(M_S, M_{S_0}) \rightarrow 0$ for all $S \in \mathcal{S}$ and $S \neq S_0$, given that the probability $f(M_{S_0})$ is bounded from zero. It turns out that to examine whether the posterior probability is consistent at the true model $M_{S_0}$ 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 $M_{S_0}$ is bounded away from zero, i.e. $f(M_{S_0}) > 0$.

**Assumption 5:** We assume $\lambda T_2 < \infty$.

**Assumption 6:** The condition
\[
\frac{(y^n)^T y^n}{\Lambda_{\min}(X_{S_0}X_{S_0}^T)} < (y^n)^T (X_S X_S^T + \lambda I)^{-1} y^n
\]
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 $M_{S_0}$ will be zero. In addition, Assumption 6 is a technical condition which ensures that the ratio $(X_{S_0}^T X_{S_0} + \lambda I_{|S_0|})/(X_S^T X_S + \lambda I_{|S|})$ is smaller than 1. The assumption will be useful in proving the convergence of the Bayes factor $BF(M_S, M_{S_0}; y^n)$.

**Theorem 5.2.** Given Assumptions 4 to 6 hold and the number of covariates $p \propto n^\alpha$, there exist some constants $0 \leq c_3 < \infty$, $\xi > 0$, $n^* > 0$ and $0 < c_{11} < \infty$ such that
\[
P(M_{S_0} | y^n) \geq 1 - c_3 \exp \left\{ - \frac{n^\alpha}{2} (n^{1-\alpha} \xi - c_{11} \log 4) \right\}
\]
for all $n > n^*$. Therefore, for $0 < \alpha < 1$, $P(M_{S_0} | y^n) \rightarrow 1$ as $n \rightarrow \infty$.

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

5.3. **Estimation consistency.** Using (5.1), the BAVA-MIO estimator can be defined as $\hat{\beta}_{BMIO} = \lim_{T_3 \rightarrow 0} \hat{\beta}^{T_3}$. Now we deal with the estimation consistency of $\hat{\beta}_{BMIO}$ under a frequentist’s framework. We will derive an asymptotic bound for the $l_2$ distance between $\hat{\beta}_{BMIO}$ and $\beta_0$ and show that the $l_2$
distance converges to 0 as \( n \to \infty \). Let \( \beta_0 \) denote the coefficient vector corresponding to the true model \( \mathcal{M}_{S_0} \). Define the expected \( l_2 \) distance between estimator \( \hat{\beta} \) and the true coefficient \( \beta_0 \) by

\[
E_Y[||\hat{\beta} - \beta_0||^2] = \int ||\hat{\beta} - \beta_0||^2 f(y|\beta_0) dy,
\]

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

**Assumption 7:** There exist some constants \( c_4 > 0 \) and \( c_5 > 0 \) such that \( \beta^2_{0,j} < c_4 \) for \( j = 1, 2, \cdots, p \), and \( \sigma^2 < c_5 \) for all \( n \).

Assumption 7 ensures that the first and second moments of parameter \( \beta_0 \) and parameter \( \sigma^2 \) 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

\[
P( ||\hat{\beta}_{BMO} - \beta_0||_2 > \xi_n ) \leq c_{13} \exp\{ - \log(n^{1-\alpha} \xi_n) \}
\]

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

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

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

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

where \( \theta \) is a parameter characterizing mean of the distribution, and \( \varphi \) is a parameter characterizing dispersion of the distribution. Under the exponential family (6.1), variable \( Y \) has properties such that \( E(Y|\theta, \varphi) = b'(\theta) \), \( Var(Y|\theta, \varphi) = b''(\theta)\varphi \). Now let \( \nu = b'(\theta) \). For a generalized linear model, there exists a link function \( \eta \) such that \( \eta(\nu) = x^T \beta \). The link function gives
a flexible connection between the mean $\nu$ and the predictor $x^T\beta$, and a valid regression can be formulated under this parametrization. In addition, $\nu$ is parametrized by $\theta$, therefore by inverse mapping, we can express $\theta$ 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 $\varphi$ in the following Bayesian hierarchical formulation. We only assign priors on regression coefficients $\beta$ and covariate indices $\gamma$. The inference concern arises from the fact that the estimation of $\varphi$ 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 $\varphi$ in the future, but at present we only focus on inference based on priors on $\beta$ and $\gamma$. Now consider the logarithm of the joint density function

$$-\log f(\beta, \gamma \mid X, y, \varphi, \lambda, \kappa) = -\sum_{i=1}^{n} \left\{ \frac{y_i \theta_i(x_i^T \beta) - b[\theta_i(x_i^T \beta)]}{\varphi} + d(y_i, \varphi) \right\}$$

$$+ \frac{\lambda}{2\varphi} \sum_{j=1}^{p} \gamma_j \beta_j^2$$

$$+ \frac{1}{2} \sum_{j=1}^{p} \gamma_j \log \left\{ \frac{2\pi \varphi (1 - \kappa)^2}{\lambda \kappa^2} \right\} + \text{const.}$$

(6.2)

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

$$\hat{\beta}^{(m+1)} = \arg \min \left\{ \sum_{i=1}^{n} y_i \theta_i(x_i^T \beta) - b[\theta_i(x_i^T \beta)] + \frac{\lambda}{2} \|\beta\|_2^2 + \rho \|\tilde{\varphi}^{(m)} \beta\|_1 \right\},$$

(6.3)

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

$$-X^TWr + \lambda \beta + g_\beta \rho \tilde{\varphi}^{(m)} = 0,$$

(6.4)

where $r = (y - \nu)\eta'(\nu)$, $W = \text{diag}\{[\eta'(\nu_1)^2]b''(\theta_1), \cdots, [\eta'(\nu_n)^2]b''(\theta_n)\}^{-1}$, $\nu = (\nu_1, \nu_2, \cdots, \nu_n)$ with $\nu_i = b'(\theta_i)$ and $g_\beta = (g_{\beta_1}, g_{\beta_2}, \cdots, g_{\beta_p})$ 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

\[
(\hat{\beta}^*(m+1)) = \arg \min_{\beta} \left\{ \frac{1}{2} ||U^{(m)}(z^{(m)} - X\beta)||_2^2 + \frac{\lambda}{2} ||\beta||_2^2 + \rho |\hat{\beta}^{(m)}||_1 \right\},
\]

(6.5)

where

\[
z^{(m)} = r^{(m)} + \eta^{(m)},
\]

\[
U^{(m)} = (W^{1/2})^{(m)}
\]

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

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

(6.6)

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

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 \( Y_i \sim \text{Bernoulli}(\zeta_i) \), where \( \zeta_i \) 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 \( \beta \), we generate \( \beta_j \) from \( \text{Normal}(0, 1) \) for \( j = 1, 2, \ldots, 10 \), and let the rest of \( 110 \beta'_{j} \)s equal to zero. We simulate covariate vector \( x_i \) i.i.d. from \( \text{MVN}(0, \Sigma_X) \). We consider three \( \Sigma_X \)'s, the same as those described in Section 4.1, to generate the covariate vectors. With \( \beta \) and \( x_i \), we simulate \( Y_i \) from \( \text{Bernoulli}(\zeta_i) \) for the cases of \( n = 100 \) and \( n = 200 \). With three different values for \( \Sigma_X \) and two different values for \( n \), we have six scenarios in the simulation study. For each scenario, we generate 100 simulation runs. In each simulation run, we apply the BAVA-MIO estimation to fit a logistic
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.

| Model       | n   | PMSE | l2-dis | S-FPR | |S| |
|-------------|-----|------|--------|-------|-----|
| BMIO-CV     | 100 | .186 (.004) | .039 (.002) | .305 (.031) | 11.39 (1.788) |
| GLM-lasso   | 100 | .173 (.003) | .044 (.004) | .680 (.012) | 21.34 (1.050) |
| BMIO-CV     | 200 | .145 (.003) | .018 (.001) | .176 (.021) | 7.54 (.549) |
| GLM-lasso   | 200 | .144 (.002) | .026 (.001) | .694 (.011) | 27.65 (1.141) |
| BMIO-CV     | 100 | .180 (.004) | .055 (.003) | .455 (.031) | 13.86 (1.841) |
| GLM-lasso   | 100 | .169 (.003) | .052 (.003) | .654 (.018) | 17.54 (.966) |
| BMIO-CV     | 200 | .157 (.003) | .027 (.002) | .327 (.025) | 8.58 (.564) |
| GLM-lasso   | 200 | .154 (.003) | .030 (.001) | .654 (.012) | 20.67 (.908) |
| BMIO-CV     | 100 | .180 (.004) | .046 (.003) | .271 (.028) | 8.28 (1.187) |
| GLM-lasso   | 100 | .170 (.003) | .047 (.003) | .668 (.016) | 19.18 (.914) |
| BMIO-CV     | 200 | .152 (.004) | .022 (.001) | .203 (.023) | 7.19 (.478) |
| GLM-lasso   | 200 | .150 (.003) | .027 (.001) | .675 (.014) | 23.50 (1.045) |

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 does.

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 been rescaled to have mean zero and variance 1, and the response variable has been centered around its mean. All estimations are based on
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) |

the rescaled covariates and centered response variable. For hyper parameters, 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 non-zero 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.
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 7129 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 $\lambda = \lambda^*\sqrt{\frac{p\log p}{n}}$ 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 [40, 32, 6, 8] 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 we did not pay much attention in this paper 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 $\lambda$ so that the estimation procedure can be carried out along the grids on $\lambda$ and $\Psi(\kappa)$. Another possible modification is to drop the prior assumption on $\sigma^2$ and treat it as a constant. In this approach the impact of $\sigma^2$ on parameter estimation can be dealt together with the tuning parameter $\Psi(\kappa)$. This approach has been adopted in Section 6 for parameter estimation in the
Table 4
Classification results for Golub's gene expression data.

| Method                        | CV-error | Test-error | # of genes |
|-------------------------------|----------|------------|------------|
| Golub et al. (1999)           | 3/38     | 4/34       | 50         |
| Elastic Net (Zou&Hastie, 2005)| 3/38     | 0/34       | 45         |
| $l_1$-pen GLM (Park&Hastie, 2007) | 1/38 | 2/34       | 23         |
| SIS-SCAD-LD (Fan&Lv, 2008)    | 0/38     | 1/34       | 16         |
| FAIR (Fan&Fan, 2008)          | 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         |

generalized linear models.

Acknowledgements. 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. Tso-Jung Yen was supported by grants NSC 97-3112-B-001-020 and NSC 98-3112-B-001-027 in the National Research Program for Genomic Medicine.

SUPPLEMENTARY MATERIAL

Supplement File: (supplemental_file.pdf). In Supplementary Material, we provide brief discussions on the log-sum function, connections with other approaches, derivation of the soft-thresholding operator, and proofs of Theorem 5.1, Theorem 5.2 and Theorem 5.3.

REFERENCES
[1] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig Selector. The Annals of Statistics 4 1705-1732.
[2] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$. The Annals of Statistics 35 2313-2351.
[3] Candès, E. J., Wakin, M. B. and Boyd, S. P. (2008). Enhancing sparsity by reweighted $l_1$ minimization. Journal of Fourier Analysis and Applications 14 877-905.
[4] Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. Journal of the Royal Statistical Society B 62 681-698.
[5] Clyde, M., Ghosh, J. and Littman, M. (2009). Bayesian adaptive sampling for variable selection.
[6] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence. The Annals of Statistics 36 2605-2637.
[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *The Journal of the American Statistical Association* 96 1348-1360.

[8] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society* 70 849-911.

[9] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33.

[10] Friedman, J., Hastie, T. H., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* 2 302-332.

[11] Genkin, A., Lewis, D. D. and Madigan, D. (2007). Large scale Bayesian logistic regression for text categorization. *Technometrics* 49 291-304.

[12] George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* 87 731-747.

[13] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal Of the American Statistical Association* 88 881-889.

[14] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7 339-374.

[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 Analysis* 5 171-188.

[17] Hans, C. (2009). Bayesian lasso regression. *Biometrika* 96 835-845.

[18] Hunter, D. R. and Lange, K. (2005). A tutorial on MM algorithms. *The American Statistician* 58 30-37.

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

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

[21] Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics* 33 1700-1752.

[22] Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* 28 1356-1378.

[23] Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis* 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. *Journal of the American Statistical Association* 103 410-423.

[25] Mazumder, R., Friedman, J. and Hastie, T. (2009). SparseNet: coordinate descent with non-convex penalties.

[26] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, New York.

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

[28] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34 1436-1462.

[29] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83 1023-1036.
[31] Park, T. and Casella, G. (2008). The Bayesian lasso. *The Journal of the American Statistical Association* **103** 681-686.

[32] Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society* **69** 659-677.

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

[34] Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* **2** 224-244.

[35] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B* **68** 49-67.

[36] Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society* **69** 143-161.

[37] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894-942.

[38] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541-2563.

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

[40] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *The Journal of the Royal Statistical Society, Series B* **67** 301-320.

[41] Zou, H. and Li, R. (2008). One-step sparse estimates in non-concave penalized likelihood models. *The Annals of Statistics* **36** 1509-1533.

[42] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37** 1733-1751.

Institute of Statistical Science
Academia Sinica
128 Academia Road, Section 2, Taipei 115
Taiwan
E-mail: tjyen@stat.sinica.edu.tw