**Electronic Journal of Statistics**
Vol. 9 (2015) 1243–1266
ISSN: 1935-7524
DOI: 10.1214/15-EJS1034

# Selecting massive variables using an iterated conditional modes/medians algorithm

**Vitara Pungpapong**

*Department of Statistics*
*Faculty of Commerce and Accountancy*
*Chulalongkorn University*
*Bangkok, Thailand*
*e-mail:* vitara@cbs.chula.ac.th

**Min Zhang and Dabao Zhang***

*Department of Statistics*
*Purdue University*
*West Lafayette, IN 47907*
*e-mail:* minzhang@stat.purdue.edu; zhangdb@stat.purdue.edu

**Abstract:** Empirical Bayes methods are designed in selecting massive variables, which may be inter-connected following certain hierarchical structures, because of three attributes: taking prior information on model parameters, allowing data-driven hyperparameter values, and free of tuning parameters. We propose an iterated conditional modes/medians (ICM/M) algorithm to implement empirical Bayes selection of massive variables, while incorporating sparsity or more complicated *a priori* information. The iterative conditional modes are employed to obtain data-driven estimates of hyperparameters, and the iterative conditional medians are used to estimate the model coefficients and therefore enable the selection of massive variables. The ICM/M algorithm is computationally fast, and can easily extend the empirical Bayes thresholding, which is adaptive to parameter sparsity, to complex data. Empirical studies suggest competitive performance of the proposed method, even in the simple case of selecting massive regression predictors.

**MSC 2010 subject classifications:** Primary 62J05; secondary 62C12, 62F07.
**Keywords and phrases:** Empirical Bayes variable selection, high dimensional data, prior, sparsity.

Received September 2014.

## Contents



*Corresponding author.

## 1. Introduction

Selecting variables in problems with a large number of predictors is a challenging yet critical problem in analyzing high-dimensional data. Because high-dimensional data are usually of relatively small sample sizes, successful variable selection demands appropriate incorporation of *a priori* information. A fundamental piece of information is that only a few of the variables are significant and should be included into the underlying models, leading to a fundamental assumption of sparsity in variable selection (Fan and Li, 2001). Many methods have been developed to take full advantage of this sparsity assumption, mostly built upon thresholding procedures (Donoho and Johnstone, 1994), see Tibshirani (1996), Fan and Li (2001), and others.

Recently, many efforts have been devoted to selecting variables from massive candidates by incorporating rich *a priori* information accumulated from historical research or practices. For example, Yuan and Lin (2006) defined group-wise norms for grouped variables. For graph-structured variables, Li and Li (2010) and Pan et al. (2010) proposed to use Laplacian matrices and $L_\gamma$ norms, respectively. Li and Zhang (2010) and Stingo et al. (2011) both employed Bayesian approaches to incorporate structural information of the variables, both formulating Ising priors.

Markov chain Monte Carlo (MCMC) algorithms have been commonly employed to develop Bayesian variable selection, see George and McCulloch (1993), Carlin and Chib (1995), Li and Zhang (2010), Stingo et al. (2011), and others. However, MCMC algorithms are computationally intensive and may be difficult to assign appropriate hyperparameters. On the other hand, penalty-based variable selection usually demands careful selection of certain tuning parameters (e.g. Fan and Li, 2001; Li and Li, 2010; Pan et al., 2010; Tibshirani, 1996; Yuan and Lin, 2006), which challenges high-dimensional data analysis. Although cross-validation has been widely suggested to choose tuning parameters, it may be infeasible in certain situations, in particular the case that many variables rarely vary. Recently, Sun and Zhang (2012) proposed the scaled sparse linear regression to attach the tuning parameter to the estimable noise level.

Empirical Bayes methods can be advantageous in high-dimensional data analysis because of no need to choose tuning parameters. They also allow incorporating *a priori* information while modeling uncertainty of such prior information using hyperparameters. For example, Johnstone and Silverman (2004) modeled the sparse normal means using a spike-and-slab prior. The mixing rate of the Dirac spike and slab is taken as a hyperparameter to achieve data-driven thresholding, and resultant empirical Bayes estimates are therefore adaptive to sparsity of the high-dimensional parameters. As demonstrated by Johnstone and Silverman (2004), this empirical Bayes method can work better than traditional thresholding estimators. One important contribution of this paper is to develop a new algorithm which allows to construct such empirical Bayes variable selection with complex data.

We propose an iterative conditional modes/medians (ICM/M) algorithm for easy implementation and fast computation of empirical Bayes variable selection (EBVS). Similar to the iterated conditional modes (Besag, 1986), iterative conditional modes are for optimization of hyperparameters and parameters other than regression coefficients. Iterative conditional medians are used to enforce variable selection. As shown in Johnstone and Silverman (2004) and Zhang et al. (2010), when mixture priors are utilized, posterior medians can lead to thresholding rules and thus help screen out small and insignificant variables. Furthermore, ICM/M makes it easy to incorporate complicated priors for the purpose of selecting variables out of massive structured candidates. Taking the Ising prior as an example (Li and Zhang, 2010), we illustrate such strength of ICM/M.

The rest of this paper is organized as follows. In the next section, we will propose the ICM/M algorithm for empirical Bayes variable selection (EBVS). We also explore to control false discovery rates (FDR) using conditional posterior probabilities. We implement the ICM/M algorithm in Section 3 for high-dimensional linear regression models, only assuming that non-zero regression coefficients are few. In Section 4, the ICM/M algorithm is shown when incorporating *a priori* information on graphical relationship between the predictors. Simulation studies are carried out in both Sections 3 and 4 to evaluate the performance of the corresponding ICM/M algorithms. An application to a real dataset from a genome-wide association study (GWAS) is presented in Section 5. We conclude this paper with a discussion in Section 6.

In the rest of this paper, the $j$-th component of a vector parameter, say $\beta$, is denoted by $\beta_j$; $\beta_{-j}$ denotes all the components of $\beta$ except the $j$-th component; and $\beta_{j:k}$ includes components of $\beta$ from $\beta_j$ to $\beta_k$. The parameter with a parenthesized superscript, say $\hat{\beta}^{(k)}$, indicates an estimate from the $k$-th iteration.

## 2. The method

### *2.1. Iterated conditional modes/medians*

Consider a general variable selection problem with a likelihood function given by,

$$\mathcal{L}(\mathbf{Y}; \mathbf{X}\beta; \phi), \tag{2.1}$$

where $\mathbf{Y}$ is a $n \times 1$ random vector, $\mathbf{X}$ is a $n \times p$ matrix containing values of $p$ variables, $\beta$ is a $p \times 1$ parameter vector with the $j$-th component $\beta_j$ representing the effects of the $j$-th variable to the model, and $\phi$ includes all other auxiliary parameters.

A typical variable selection task is to identify non-zero components in $\beta$, that is, to select important variables out of the $p$ candidates. For convenience, define $\tau_j = I\{\beta_j \neq 0\}$, which indicates whether the $j$-th variable should be selected into the model. Further denote $\tau = (\tau_1, \tau_2, \ldots, \tau_p)^t$. Here we consider an empirical Bayes variable selection, which assumes priors,

$$\begin{cases} \beta \sim \pi(\beta|\tau, \psi_1) \times \pi(\tau|\psi_2), \\ \phi \sim \pi(\phi|\psi_3), \end{cases} \tag{2.2}$$

where $\psi = (\psi_1^t, \psi_2^t, \psi_3^t)^t$ includes all hyperparameters.

To avoid high-dimensional integrals, here we cycle through coordinates to obtain the estimate of each component of $(\beta, \phi, \psi)$ iteratively,

$$\begin{cases} \hat{\beta}_j = \hat{\beta}_j(\hat{\beta}_{-j}, \hat{\phi}, \hat{\psi}), \\ \hat{\phi}_j = \hat{\phi}_j(\hat{\beta}, \hat{\phi}_{-j}, \hat{\psi}), \\ \hat{\psi}_j = \hat{\psi}_j(\hat{\beta}, \hat{\phi}, \hat{\psi}_{-j}). \end{cases} \tag{2.3}$$

Indeed, with properly chosen priors of $\phi$ and $\psi$, both $\hat{\phi}_j = \hat{\phi}_j(\hat{\beta}, \hat{\phi}_{-j}, \hat{\psi}, \mathbf{Y}, \mathbf{X})$ and $\hat{\psi}_j = \hat{\psi}_j(\hat{\beta}, \hat{\phi}, \hat{\psi}_{-j}, \mathbf{Y}, \mathbf{X})$ can be obtained by maximizing the fully conditional posterior, i.e.,

$$\begin{cases} \hat{\phi}_j = \hat{\phi}_j(\hat{\beta}, \hat{\phi}_{-j}, \hat{\psi}) = \text{mode}(\phi_j|\mathbf{Y}, \mathbf{X}, \hat{\beta}, \hat{\phi}_{-j}, \hat{\psi}), \\ \hat{\psi}_j = \hat{\psi}_j(\hat{\beta}, \hat{\phi}, \hat{\psi}_{-j}) = \text{mode}(\psi_j|\mathbf{Y}, \mathbf{X}, \hat{\beta}, \hat{\phi}, \hat{\psi}_{-j}). \end{cases} \tag{2.4}$$

When each $\hat{\beta}_j$ is also obtained by maximizing its fully conditional posterior, it suggests the iterated conditional modes (ICM) algorithm by Besag (1986). However, calculation of conditional mode for $\hat{\beta}_j$ is either infeasible or practically undesirable (due to lack of variable selection function). Indeed, Bayesian or empirical Bayes variable selection for high-dimensional data usually follows a spike-and-slab prior on each $\beta_j$ (e.g. Ishwaran and Rao, 2005; Mitchell and Beauchamp, 1988), and it induces a spike-and-slab posterior for each $\beta_j$. With a Dirac spike, it is infeasible to obtain the mode of such a spike-and-slab posterior. However, its median can be zero and allows to select the median probability model as suggested by Barbieri and Berger (2004). Henceforth, following Johnstone and Silverman (2004), we construct $\hat{\beta}_j = \hat{\beta}_j(\hat{\beta}_{-j}, \hat{\phi}, \hat{\psi}, \mathbf{Y}, \mathbf{X})$ as median of the fully conditional posterior, i.e.,

$$\hat{\beta}_j = \hat{\beta}_j(\hat{\beta}_{-j}, \hat{\phi}, \hat{\psi}) = \text{median}(\beta_j|\mathbf{Y}, \mathbf{X}, \hat{\beta}_{-j}, \hat{\phi}, \hat{\psi}). \tag{2.5}$$

With the iterative conditional median for $\beta_j$, and conditional modes for $\phi_j$ and $\psi_j$ respectively, for Bayesian update of a component conditional on all other components, we hereafter propose the iterated conditional medians/modes

(ICM/M) algorithm for implementing the empirical Bayes variable selection. As shown later, the ICM/M algorithm allows an easy extension of the (generalized) empirical Bayes thresholding methods by Johnstone and Silverman (2004) and Zhang et al. (2010) to dependent data. Obviously, with a consistent initial point of $(\hat{\beta}, \hat{\phi}, \hat{\psi})$, the cycling Bayesian updates of this algorithm lead to a well-established estimate $(\hat{\beta}, \hat{\phi}, \hat{\psi})$.

### *2.2. Evaluation of variable importance*

When proposing a statistical model, we are primarily interested in evaluating the importance of variables besides its predictive ability. For example, the objective of high-dimensional data analysis is to identify a list of $J$ predictors that are most important or significant among $p$ predictors. This is a common practice in biomedical research using high-throughput biotechnologies, ranking all markers and selecting a short list of candidates for follow-up studies.

In the Bayesian approach, inference on the importance of each variable can be done through its marginal posterior probability $P(\beta_j \neq 0|\mathbf{Y}, \mathbf{X})$. However, this quantity involves high-dimensional integrals which is difficult to calculate even in the case of moderate $p$. Furthermore, the marginal posterior probability may not be meaningful when the predictors are highly correlated (which usually occurs in a large $p$ small $n$ data set). For example, suppose predictors $X_1$ and $X_2$ are linearly dependent and both predictors are associated with a response variable. The marginal posterior probability of $X_1$ being included in the model might be very high and dominates the marginal posterior probability of $X_2$ being included in the model.

We propose a local posterior probability to evaluate the importance of a variable. That is, conditional on the optimal point $\{\hat{\beta}_j, \hat{\phi}, \hat{\psi}\}$ obtained from empirical Bayes variable selection through ICM/M algorithm, the importance of a variable is evaluated by its full conditional posterior probability,

$$\zeta_j = P(\beta_j \neq 0|\mathbf{Y}, \mathbf{X}, \hat{\beta}_{-j}, \hat{\phi}, \hat{\psi}). \tag{2.6}$$

Such a probability has a closed form which can be easily computed. We will show later in simulation studies that the local posterior probability is a good indicator to quantify the importance of variables.

Another challenging question is how large the list of important predictors should be. In many papers in the literature, see Brenner et al. (2001) and Syed and Hecht (1997) for example, the numbers of important variables reported are arbitrary. For instance, some laboratories may be interested in the top ten genes. Typically, however, there is an interest to create the list so that type-I and type-II errors are controlled (Dudoit et al., 2003). False discovery rate (FDR) control is widely used in high-dimensional data analysis since it is less conservative and has more power than controlling the familywise error rate (Benjamini and Hochberg, 1995).

With the local posterior probability $\zeta$ and assumption that true $\beta$ is known, we can report a list containing predictors having the posterior probability greater

than some bound $\kappa$, $0 \leq \kappa < 1$. Given the data, true FDR can be computed as

$$FDR(\kappa) = \sum_{j=1}^{p} I\{\beta_j = 0, \zeta_j > \kappa\} \Big/ \sum_{j=1}^{p} I\{\zeta_j > \kappa\}. \tag{2.7}$$

Newton et al. (2004) proposed the expected FDR given the data in Bayesian scheme as

$$\widehat{FDR}(\kappa) = \sum_{j=1}^{p} (1 - \zeta_j) I\{\zeta_j > \kappa\} \Big/ \sum_{j=1}^{p} I\{\zeta_j > \kappa\}. \tag{2.8}$$

Therefore we can select predictors to report by controlling $\widehat{FDR}(\kappa)$ at a desired level.

## 3. Selection of sparse variables

Here we consider the empirical Bayes variable selection for the following regression model with high dimensional data,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n). \tag{3.1}$$

Further assume that the response is centered and the predictors are standardized, that is, $\mathbf{Y}^t \mathbf{1}_n = 0$, $\mathbf{X}^t \mathbf{1}_n = \mathbf{0}_p$, and

$$\mathbf{X}_j^t \mathbf{X}_j = n - 1, \quad j = 1, \ldots, p,$$

where $\mathbf{X}_j$ is the $j$-th column of $\mathbf{X}$, i.e., $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$.

Let $\tilde{\mathbf{Y}}_j = \mathbf{Y} - \mathbf{X}\beta + \mathbf{X}_j\beta_j$. Assuming all model parameters except $\beta_j$ are known, $\beta_j$ has a sufficient statistic

$$\frac{1}{n-1} \mathbf{X}_j^t \tilde{\mathbf{Y}}_j \sim N\left(\beta_j, \frac{1}{n-1}\sigma^2\right). \tag{3.2}$$

To capture the sparsity of regression coefficients, we put an independent prior on each scaled $\beta_j$ as follows,

$$\beta_j | \sigma \overset{iid}{\sim} (1 - \omega)\delta_0(\cdot) + \omega\gamma(\cdot|\sigma), \tag{3.3}$$

where $\delta_0(\cdot)$ is a Dirac delta function at zero, $\gamma(\cdot|\sigma)$ is assumed to be a probability density function. This mixture prior implies that $\beta_j$ is zero with probability $(1 - \omega)$ and is drawn from the nonzero part of prior, $\gamma(\cdot|\sigma)$, with probability $\omega$. As suggested by Johnstone and Silverman (2004), a heavy-tailed prior such as Laplace distribution is a good choice for $\gamma(\cdot|\sigma)$, that is,

$$\gamma(\beta_j|\sigma) = \frac{\alpha\sqrt{n-1}}{2\sigma} \exp\left(-\frac{\alpha\sqrt{n-1}}{\sigma}|\beta_j|\right), \tag{3.4}$$

where $\alpha > 0$ is a scale parameter. We take Jeffreys' prior on $\sigma$ as $\pi(\sigma) \propto 1/\sigma$ (Jeffreys, 1946).

Note that there is a connection of using Laplace priors and the lasso. Indeed, setting $\omega = 1$ in (2.4) and (3.3) leads to a lasso estimate with $\alpha$ related to a tuning parameter in the lasso, see details in Tibshirani (1996). Our empirical

Bayes variable selection allows a data-driven optimal choice of $\omega$. Indeed, a data-driven optimal $\alpha$ can also be obtained through the conditional mode suggested by (2.4), which avoids the issue brought by a tuning parameter to lasso (while lasso usually relies on cross validation to choose an optimal tuning parameter). Johnstone and Silverman (2004) also suggested a default value $\alpha = 0.5$, which in general works well.

### *3.1. The algorithm*

Here we implement the ICM/M algorithm described in (2.4) and (2.5). Note that $\phi = \sigma$, and $\psi = (\omega, \alpha)$ or $\psi = \omega$ depending on whether $\alpha$ is fixed. Throughout this paper, we fix $\alpha = 0.5$ as suggested by Johnstone and Silverman (2004).

To obtain $\hat{\beta}_j^{(k+1)} = \text{median}(\beta_j|\mathbf{Y}, \mathbf{X}, \hat{\beta}_{1:(j-1)}^{(k+1)}, \hat{\beta}_{(j+1):p}^{(k)}, \hat{\sigma}^{(k)}, \hat{\omega}^{(k)})$, we notice the sufficient statistic of $\beta_j$ in (3.2) and it is therefore easy to calculate $\hat{\beta}_j^{(k+1)}$ as stated below. Indeed, $\hat{\beta}_j^{(k+1)}$ is an empirical Bayes thresholding estimator as shown in Johnstone and Silverman (2004).

**Proposition 3.1.** *With pre-specified values of $\sigma$ and $\beta_{-j}$, $\frac{1}{n-1}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j$ is a sufficient statistic for $\beta_j$ w.r.t the model (3.1). Furthermore, the iterative conditional median of $\beta_j$ in the ICM/M algorithm can be constructed as the posterior median of $\beta_j$ in the following Bayesian analysis,*

$$\begin{cases} \frac{1}{\sigma\sqrt{n-1}}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j|\beta_j \sim N\left(\frac{\sqrt{n-1}}{\sigma}\beta_j, 1\right), \\ \beta_j \sim (1-\omega)\delta_0(\beta_j) + \omega\frac{\sqrt{n-1}}{4\sigma}\exp\left(-\frac{\sqrt{n-1}}{2\sigma}|\beta_j|\right). \end{cases}$$

The conditional mode $\hat{\sigma}^{(k+1)} = \text{mode}(\sigma|\mathbf{Y}, \mathbf{X}, \hat{\beta}^{(k+1)}, \hat{\omega}^{(k)})$ has an explicit solution,

$$\hat{\sigma}^{(k+1)} = \frac{1}{4d}\left(c + \sqrt{c^2 + 16d||\mathbf{Y} - \mathbf{X}\hat{\beta}^{(k+1)}||^2}\right),$$

where $c = \sqrt{n-1}\|\hat{\beta}^{(k+1)}\|_1$, and $d = n + \|\hat{\beta}^{(k+1)}\|_0 + 1$. Furthermore, the conditional mode $\hat{\omega}^{(k+1)} = \text{mode}(\omega|\mathbf{Y}, \mathbf{X}, \hat{\beta}^{(k+1)}, \hat{\sigma}^{(k+1)})$ can be easily calculated as

$$\hat{\omega}^{(k+1)} = \|\hat{\beta}^{(k+1)}\|_0/p.$$

### *3.2. Simulation studies*

To evaluate the performance of our proposed empirical Bayes variable selection (EBVS) via ICM/M algorithm, we simulated data from model (3.1) with large $p$ small $n$, i.e., $p = 1{,}000$ and $n = 100$. There are a total of 20 non-zero regression coefficients which are $\beta_1 = \cdots = \beta_{10} = 2$ and $\beta_{101} = \cdots = \beta_{110} = 1$. The error standard deviation $\sigma$ is set to one. The predictors are partitioned into ten blocks, each including 100 predictors which are serially correlated at the same

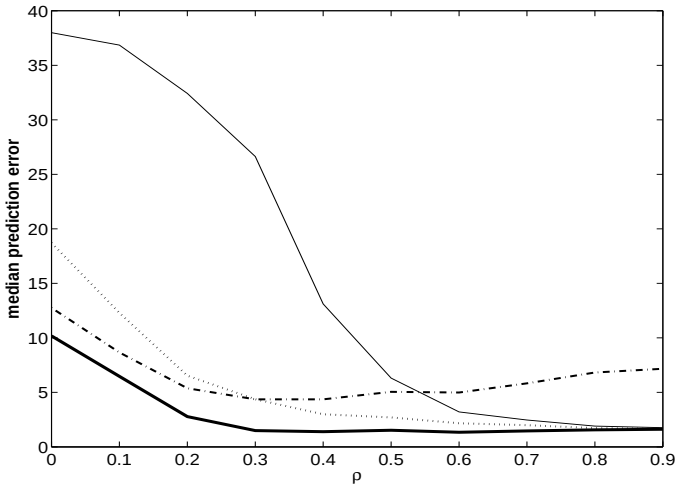


FIG 1. *Comparison of median prediction errors of lasso (dotted), adaptive lasso (dash-dotted), scaled lasso (thin solid), and EBVS (thick solid) by averaging over 100 datasets simulated for each $\rho$ in Section 3.2.*

level of correlation coefficient $\rho$. We simulated 100 datasets for $\rho$ taking values in $\{0, 0.1, 0.2, \ldots, 0.9\}$ respectively.

EBVS was compared with two popular approaches, i.e., the lasso by Tibshirani (1996), and the adaptive lasso by Zou (2006). The scaled lasso by Sun and Zhang (2012) was also applied to the simulated datasets. Ten-fold cross-validation was used to choose optimal tuning parameters for lasso and adaptive lasso respectively. The median values of prediction error, false positive, and false negative rates were reported for each approach based on the 100 simulated datasets.

As shown in Figure 1, EBVS performs much better than the other three methods in terms of prediction error. In particular, when $\rho \geq 0.3$, EBVS consistently reported median prediction error approximately at 1.5. In comparison of lasso and adaptive lasso, adaptive lasso has smaller prediction error when $\rho < 0.3$; but lasso has smaller prediction error when $\rho > 0.3$.

It is known that lasso can inconsistently select variables under certain conditions, and adaptive lasso was proposed for solving this issue (Zou, 2006). Figure 2 showed that lasso has very high false positive rates (more than 50%), and adaptive lasso significantly reduces the false positive rates especially when $\rho \geq 0.2$. Indeed, lasso has much larger false positive rates than all other methods. It is interesting to observe that EBVS has zero false positive rates except in the case that $\rho = 0.5$ and $\rho = 0.9$. All methods have very low false negative rates.

Recently, Meinshausen et al. (2009) proposed a multi-sample-split method to construct p-values for high-dimensional regressions, especially in the case that the number of predictors is larger than the sample size. Here we applied this method, as well as EBVS, to each simulated dataset with a total of 50 sample-splits, and compared its performance with that of $\zeta_i$ defined in (2.6).

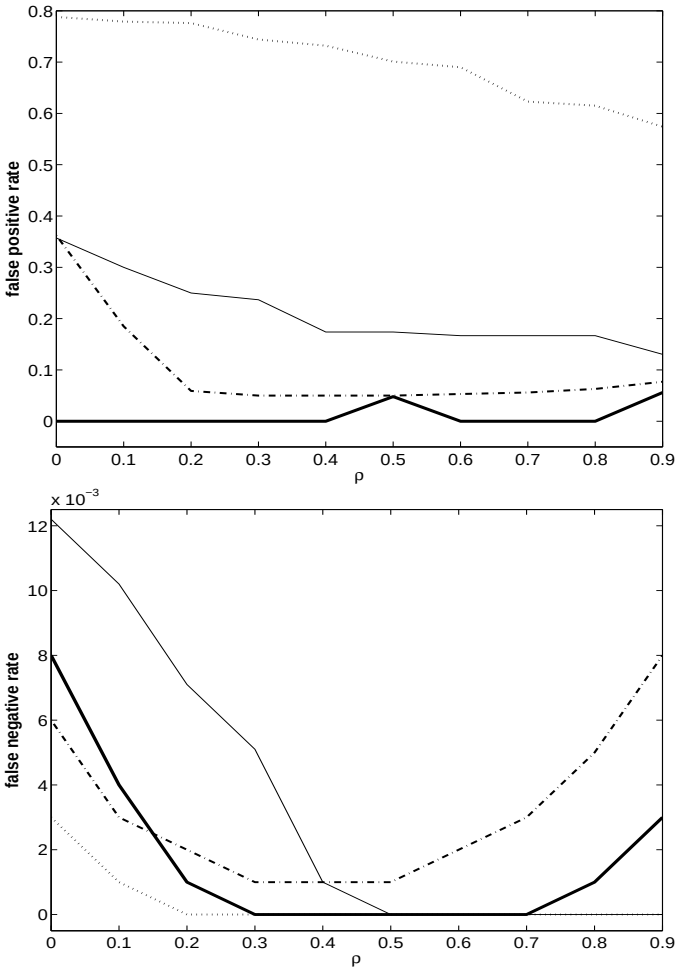


FIG 2. *Comparison of false positive rates (top) and false negative rates (bottom). Averaging over 100 datasets simulated for each ρ in Section 3.2, the false positive/negative rates were calculated for lasso (dotted), adaptive lasso (dash-dotted), scaled lasso (thin solid), and EBVS (thick solid).*

For each predictor, Figure 3 plotted the median of $-\log_{10}(1-\zeta_i)$, truncated at 10, against the median of $-\log_{10}$(p-value) across 100 datasets simulated from the regression model with $\rho = 0.5$ and $\rho = 0.9$ respectively. For either model, $\zeta_i$ can clearly distinguish true positives (i.e., predictors with $\tau_i \neq 0$) from true negatives (i.e., predictors with $\tau_i = 0$). However, as shown in Figure 3.b where $\rho = 0.9$, there is no clear cutoff of p-values to distinguish between true positives and true negatives. Here we also observed that $FDR(\kappa)$ can be well approximated by $\widehat{FDR}(\kappa)$ (results are not shown), with both dropped sharply to zero for $\kappa > 0.05$. We therefore can select $\kappa$ to threshold $\zeta_i$ for the purpose of controlling FDR.

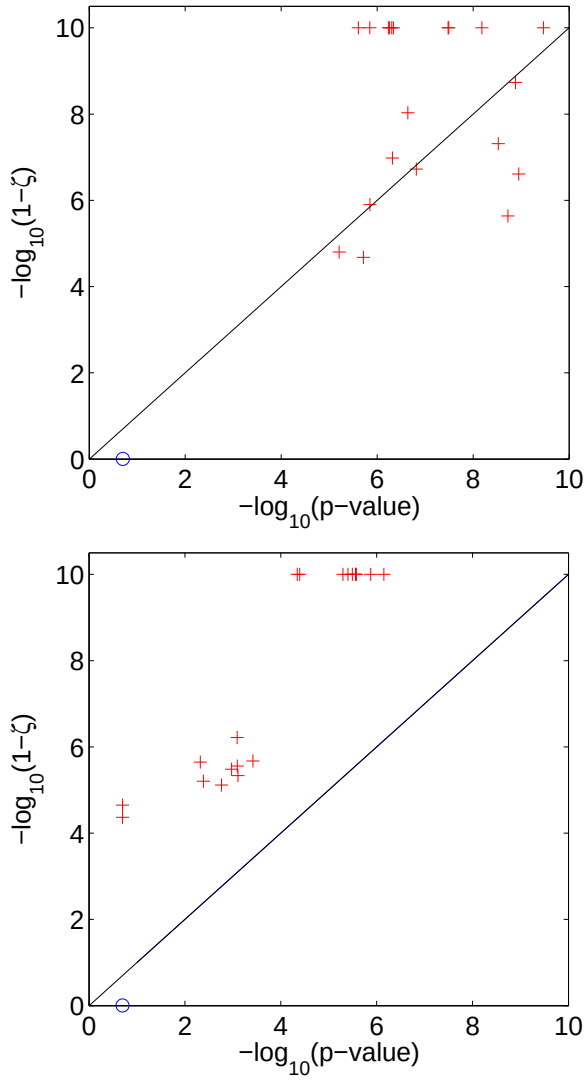


FIG 3. *Comparison of the local posterior probabilities (with* $-\log_{10}(1-\zeta)$ *truncated at* 10*) and p-values in evaluating variable importance by EBVS, with* $\rho = 0.5$ *(top) and* $\rho = 0.9$ *(bottom). Each plot is based on 100 datasets simulated in Section 3.2. True positives are indicated by crosses and true negatives are indicated by circles.*

## 4. Selection of structured variables

When the information of structural relationships among predictors is available, it is unreasonable to assume an independent prior on each $\beta_j, j = 1, \ldots, p$ as described in the previous section. Instead, we introduce an indicator variable $\tau_j = I\{\beta_j \neq 0\}$. Then, the prior distribution of $\beta$ is set to be dependent on

$\tau = (\tau_1, \ldots, \tau_p)^T$. Specifically, given $\tau_j$, $\beta_j$ has the mixture distribution

$$\beta_j|\tau_j \sim (1-\tau_j)\delta_0(\beta_j) + \tau_j\gamma(\beta_j), \tag{4.1}$$

where $\gamma(\cdot)$ is the Laplace density with the scale parameter $\alpha$.

The relationship among predictors can be represented by an undirected graph $G = (V, E)$ comprising a set $V$ of vertices and a set $E$ of edges. In this case, each node is associated with a binary valued random variable $\tau_j \in \{0, 1\}$ and there is an edge between two nodes if two covariates are correlated. The following Ising model (Onsager, 1943) is employed to model the *a priori* information on $\tau$,

$$P(\tau) = \frac{1}{Z(a,b)} \exp\left\{ a\sum_i \tau_i + b \sum_{<i,j>\in E} \tau_i\tau_j \right\}, \tag{4.2}$$

where $a$ and $b$ are two parameters, and

$$Z(a,b) = \sum_{\tau\in\{0,1\}^p} \exp\left\{ a\sum_i \tau_i + b \sum_{<i,j>\in E} \tau_i\tau_j \right\}.$$

The parameter $b$ corresponds to the "energies" associated with interactions between nearest neighboring nodes. When $b > 0$, the interaction is called *ferromagnetic*, i.e., neighboring $\tau_i$ and $\tau_j$ tend to have the same value. When $b < 0$, the interaction is called *antiferromagnetic*, i.e., neighboring $\tau_i$ and $\tau_j$ tend to have different values. When $b = 0$, there is no interaction, and the prior gets back to independent and identical Bernoulli distribution. The value of $a + b$ indicates the preferred value of each $\tau_i$. That is, if $a + b > 0$, $\tau_i$ tends to be one; if $a + b < 0$, $\tau_i$ tends to be zero.

### *4.1. The algorithm*

Next, we describe the ICM/M algorithm to develop empirical Bayes variable selection with Ising prior (abbreviated as EBVS$_i$) to incorporate the structure of predictors in modeling process. We assume the Ising prior as homogeneous Boltzmann model, but the algorithm can be extended to more general priors. With $\alpha = 0.5$, the ICM/M algorithm described in (2.4) and (2.5) can be proceeded with $\phi = \sigma$ and $\psi = (\omega, a, b)$.

For the hyperparameters $a$ and $b$, we will calculate the conditional mode of $(a, b)$ simultaneously. Conceptually, we want $(\hat{a}^{(k+1)}, \hat{b}^{(k+1)})$ maximizing the prior likelihood $P(\tau)$ in (4.2). However, it requires to compute $Z(a, b)$ by summing up $p$-dimensional space of $\tau$, which demands intensive computation especially for a large $p$. Many methods have been proposed for approximate calculation, see Geyer (1991), Geyer and Thompson (1992), Zhou and Schmidler (2009) and others. Here we will consider the composite likelihood approach (Varin et al., 2011) which is widely used when the actual likelihood is not easy to compute. In particular, $(\hat{a}^{(k+1)}, \hat{b}^{(k+1)})$ will be obtained by maximizing a pseudo-likelihood

function, a special type of composite conditional likelihood and a natural choice for a graphical model (Besag, 1975).

With the Ising prior on $\tau^{(k)}$, the pseudo-likelihood of $(a, b)$ is as follows,

$$L_p(a,b) = \prod_{i=1}^{p} P(\tau_i^{(k)} | \tau_{-j}^{(k)}, a, b) = \prod_{i=1}^{p} \frac{\exp\left\{\tau_i^{(k)}(a + b\sum_{<i,j>\in E} \tau_j^{(k)})\right\}}{1 + \exp\left\{a + b\sum_{<i,j>\in E} \tau_j^{(k)}\right\}}.$$

The surface of such a pseudo-likelihood is much smoother than the joint likelihood and therefore easy to maximize (Liang and Yu, 2003). The resultant estimator $(\hat{a}^{(k+1)}, \hat{b}^{(k+1)})$ by maximizing $L_p(a, b)$ is biased for a finite sample size, but it is asymptotically unbiased and consistent (Guyon and Kunsch, 1992; Mase, 2000; Varin et al., 2011). The implementation of the pseudo-likelihood method is fast and straightforward which is feasible for large scale graphs. Indeed, $\hat{a}^{(k+1)}$ and $\hat{b}^{(k+1)}$ are the logistic regression coefficients when the binary variable $\hat{\tau}_i^{(k)}$ is regressed on $\sum_{<i,j>\in E} \hat{\tau}_j^{(k)}$ for $i = 1, \ldots, p$.

As shown in the previous sections, the conditional median $\hat{\beta}_j^{(k+1)}$ can be constructed on the basis of the following proposition.

**Proposition 4.1.** *With pre-specified values of $\sigma$, $a$, $b$, and $\beta_{-j}$, $\frac{1}{n-1}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j$ is a sufficient statistic for $\beta_j$ w.r.t the model (3.1). Furthermore, the iterative conditional median of $\beta_j$ in the ICM/M algorithm can be constructed as the posterior median of $\beta_j$ in the following Bayesian analysis,*

$$\begin{cases} \frac{1}{\sigma\sqrt{n-1}}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j|\beta_j \sim N\left(\frac{\sqrt{n-1}}{\sigma}\beta_j, 1\right), \\ \beta_j \sim (1-\varpi_j)\delta_0(\beta_j) + \varpi_j\frac{\sqrt{n-1}}{4\sigma}\exp\left(-\frac{\sqrt{n-1}}{2\sigma}|\beta_j|\right), \end{cases}$$

*where the probability $\varpi_j$ is specified as follows,*

$$\varpi_j^{-1} = 1 + \exp\left\{-a - b\sum_{k:<j,k>\in E} \tau_k\right\}.$$

The conditional mode $\hat{\sigma}^{(k+1)} = \text{mode}(\sigma|\mathbf{Y}, \mathbf{X}, \hat{\beta}^{(k+1)}, \hat{\omega}^{(k)})$ has an explicit solution,

$$\hat{\sigma}^{(k+1)} = \frac{1}{4d}\left(c + \sqrt{c^2 + 16d||\mathbf{Y} - \mathbf{X}\hat{\beta}^{(k+1)}||^2}\right),$$

where $c = \sqrt{n-1}\|\hat{\beta}^{(k+1)}\|_1$, and $d = n + \|\hat{\beta}^{(k+1)}\|_0 + 1$.

### *4.2. Simulation studies*

Here we simulated large $p$ small $n$ datasets from model (3.1) with structured predictors, i.e., the values of $\beta_j$ depend on correlated $\tau_j$. We here consider two different correlation structures of $\tau_i$. Both EBVS and EBVS$_i$ were applied to each simulated dataset, and they were compared with three other methods, i.e., lasso, adaptive lasso, and scaled lasso.

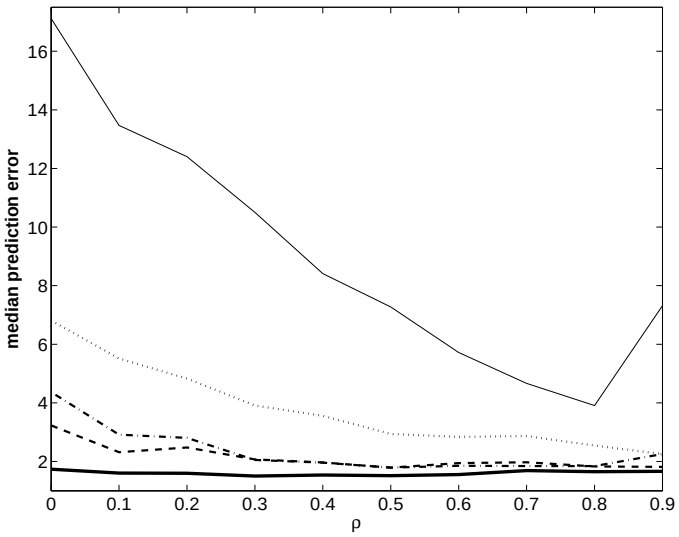


FIG 4. *Comparison of median prediction errors of lasso (dotted), adaptive lasso (dash-dotted), scaled lasso (thin solid), EBVS (dashed), and EBVS$_i$ (thick solid) by averaging over 100 datasets simulated for each $\rho$ in Case I of Section 4.2.*

**Case I. Markov Chain** For each $j = 1, \ldots, p$, $\beta_j = 0$ if $\tau_j = 0$; and if $\tau_j = 1$, $\beta_j$ is independently sampled from a uniform distribution on $[0.3, 2]$. The indicator variables $\tau_1, \ldots, \tau_p$ form a Markov chain with the transition probabilities specified as follows,

$$P(\tau_{j+1} = 0|\tau_j = 0) = 1 - P(\tau_{j+1} = 1|\tau_j = 0) = 0.99;$$
$$P(\tau_{j+1} = 0|\tau_j = 1) = 1 - P(\tau_{j+1} = 1|\tau_j = 1) = 0.5.$$

The first indicator variable $\tau_1$ is sampled from Bernouli(0.5). The error variance is fixed at one. For each individual, its predictors were simulated from $AR(1)$ with correlation coefficient $\rho$ ranging from 0 to 0.9 with step 0.1.

The median prediction error rates of all methods are shown in Figure 4. EBVS performed slightly better than adaptive lasso, and both performed much better than lasso and scaled lasso. Lasso, adaptive lasso, scaled lasso, and EBVS all presented varying prediction error rates when $\rho$ goes from 0 to 0.9. However, the prediction error rates of EBVS$_i$ are rather stable for varying values of $\rho$, and are much smaller than those of the other four methods.

Shown in Figure 5 are the false positive rates and false negative rates of different methods. Not surprisingly, lasso has false positive rates over 70%, much higher than that of other methods. Adaptive lasso significantly reduces the false positive rates, which is still more than 10%. On the other hand, the false positive rates of both EBVS and EBVS$_i$ are less than 10%. Indeed, EBVS reported false positive rates at zero for different values of $\rho$; and EBVS$_i$ reported false positive rates at zero when $\rho < 0.6$, and 0.1 when $\rho \geq 0.6$. However, EBVS$_i$ reported false negative rates less than EBVS. Therefore, EBVS tends to select correct

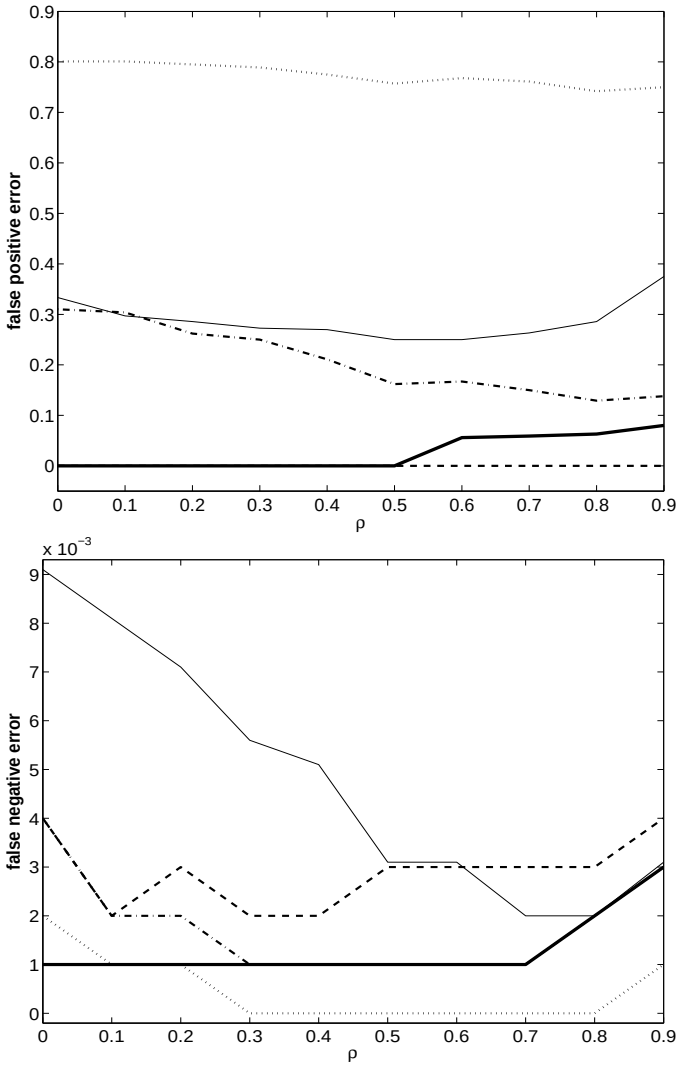


FIG 5. *Comparison of false positive rates (top) and false negative rates (bottom). Averaging over 100 datasets simulated for each $\rho$ in Case I of Section 4.2, the false positive/negative rates were calculated for lasso (dotted), adaptive lasso (dash-dotted), scaled lasso (thin solid), EBVS (dashed), and EBVS$_i$ (thick solid).*

true positives by including fewer true positives in the final model than the model obtained by EBVS$_i$. We then conjecture that, when covariates are highly correlated, EBVS$_i$ tends to select more variables into the model. In particular, if one covariate is selected into the model, EBVS$_i$ tends to include its highly correlated neighboring predictors into the model.

Figure 6 shows $FDR(\kappa)$ and $\widehat{FDR}(\kappa)$ of EBVS$_i$ for the models with $\rho = 0.5$ and $\rho = 0.9$ respectively (we also observed that $FDR(\kappa)$ of EBVS is similar to

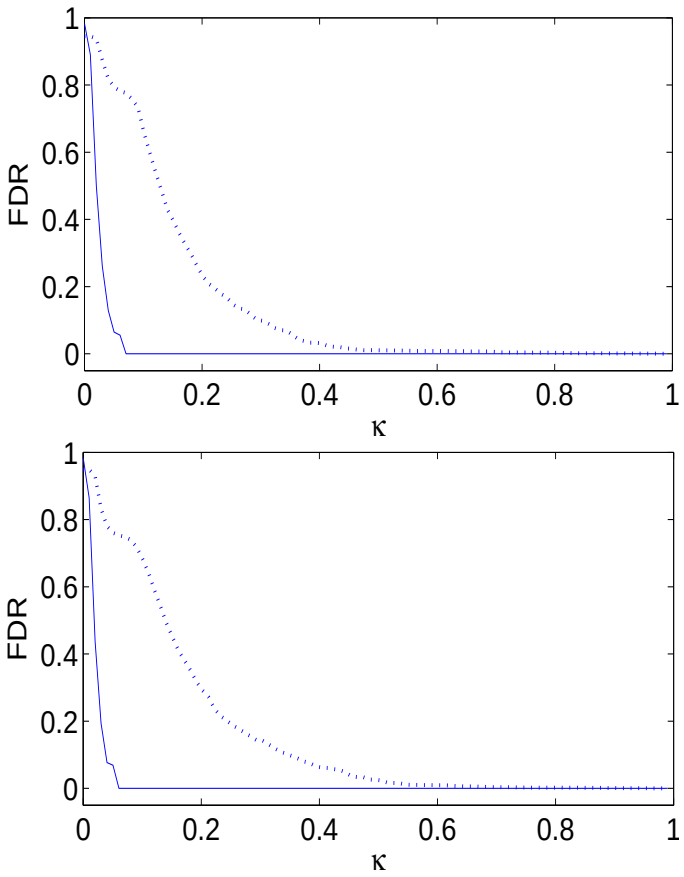


FIG 6. *Plots of median true FDR (solid) and estimated FDR (dotted) versus* $\kappa$ *based on the results of applying EBVS*$_i$ *to 100 data simulated for Case I in Section 4.2, with* $\rho = 0.5$ *(top) and* $\rho = 0.9$ *(bottom) respectively.*

that of EBVS$_i$, results are not shown). Overall, the estimate $\widehat{FDR}(\kappa)$ dominates $FDR(\kappa)$, i.e., the true FDR. Therefore, we will be conservative in selecting variables when controlling FDR using $\widehat{FDR}(\kappa)$. For example, if one would like to list important predictors while controlling FDR at 0.1 for the model with $\rho = 0.9$, $\kappa$ should be set around 0.1 based on $FDR(\kappa)$. However, one can set $\kappa$ around 0.4 based on $\widehat{FDR}(\kappa)$, which suggests a true FDR as low as zero.

Plotted in Figure 7 are the p-values calculated using the multi-sample-split method (Meinshausen et al., 2009) against $\zeta_j$ for each predictor. For both EBVS and EBVS$_i$, $\zeta_j$ quantified variable importance better than p-values in terms of distinguishing true positives from true negatives. Overall, EBVS$_i$ outperforms EBVS since it provides larger values of $\zeta$ for true positives, while both EBVS and EBVS$_i$ keep true negatives with $\zeta_j$ close to zero. Indeed, EBVS produced $\zeta_j$ close to 0 for several true positives while EBVS$_i$ produced larger values of $\zeta_j$ for

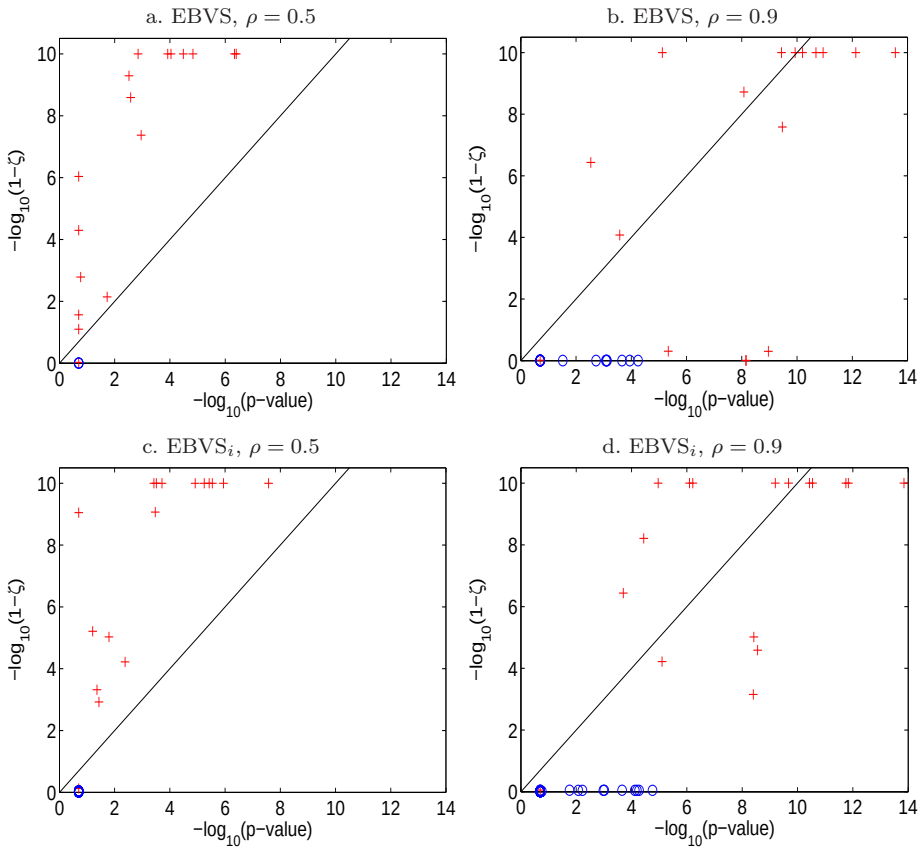


FIG 7. *Comparison of local posterior probabilities (with* $-\log_{10}(1-\zeta)$ *truncated at* 10*) and p-values in evaluating variable importance by EBVS and EBVS$_i$. Each plot is based on 100 datasets simulated for Case I in Section 4.2. True positives are indicated by crosses and true negatives are indicated by circles.*

these true positives. We then summarize empirically that, by incorporating *a priori* information, EBVS$_i$ has more power to detect true positives than EBVS.

**Case II. Pathway Information** To mimic a real genome-wide association study (GWAS), we took values of some single nucleotide polymorphisms (SNPs) in the Framingham dataset (Cupples et al., 2007) to generate $\mathbf{X}$ in model (3.1). Specifically, 24 human regulatory pathways were retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG) database, and involved 1,502 genes. For each gene involved in these pathways, at most two SNPs listed in the Framingham dataset were randomly selected out of those SNPs residing in the genetic region. If no SNP could be found within the genetic region, a nearest neighboring SNP would be identified. A total of 1,782 SNPs were selected. We first identified 952 unrelated individuals out of the Framingham dataset, and used them to generate predictor values of the training dataset. For the rest of the

TABLE 1
*Results of Simulation Studies with Pathway Information (Case II)*

| Method | Prediction Error (s.e.) | False Positive (s.e.) | False Negative (s.e.) |
|---|---|---|---|
| LASSO | 30.6928(.4050) | .6905(.0004) | .0204(.0004) |
| $\text{LASSO}_a$ | 206.1994(.5726) | .0744(.0017) | .1266(.0002) |
| $\text{LASSO}_s^*$ | 368.6464(6.1308) | .1290(.0077) | .1475(.0012) |
| EBVS | 95.3686(1.8820) | .0118(.0010) | .0970(.0008) |
| $\text{EBVS}_i$ | 21.7731(.2320) | .0308(.0015) | .0394(.0003) |

* The results of the scaled lasso excluded seven datasets. Applying the scaled lasso to these seven datasets reported the median prediction error at $2.45 \times 10^{10}$, false positive rate at .7059, and false negative rate at .0043.

Framingham dataset, we identified 653 unrelated individuals to generate predictor values of the test dataset. Five pathways were assumed to be associated with the phenotype $Y$. That is, all 311 SNPs involved in these five pathways were assumed to have nonzero regression coefficients, which were randomly sampled from a uniform distribution over $[0.5, 3]$. With the error variance at five, a total of 100 datasets were simulated.

As shown in Table 1, lasso has relatively low prediction error. However, its median false positive rate is as high as 69%, much higher than others. Adaptive lasso ($\text{LASSO}_a$), on the other hand, has very large prediction error, but its false positive rate is much smaller than lasso. EBVS presented the lowest false positive rate among all the methods, and its false negative rate is also smaller than that of adaptive lasso. Indeed, with initial values obtained from lasso, EBVS reduces the false positive rate from lasso by more than 98%. By incorporating the pathway information using an Ising prior on $\tau$, $\text{EBVS}_i$ reported the lowest prediction error. Furthermore, $\text{EBVS}_i$ compromised between lasso, adaptive lasso, and EBVS to balance well between the false positive rate and false negative rate. Scaled lasso ($\text{LASSO}_s$) performed unstably in analyzing our simulated datasets, and it selected more than 800 positives in seven of the simulated datasets.

## 5. Real data analysis

The empirical Bayes variable selection using ICM/M algorithm was applied to the Framingham dataset (Cupples et al., 2007) to find SNPs associated with vitamin D level. The SNPs of the dataset were preprocessed following common criteria of GWAS, that is, both missingness per individual and missingness per SNP are less than 10%; minor allele frequency (MAF) is no less than 5%; and the significance level of Hardy-Weinberg test on each SNP is 0.001. It resulted in a total of 370,773 SNPs, and 84,834 of them resided in 2,167 genetic regions involving 112 pathways relevant to vitamin D level. We pre-screened SNPs by selecting those having p-values of univariate tests smaller than 0.1, and ended with 7,824 SNPs for the following analysis. As in Section 4.2, a training dataset and a test dataset were constructed with 952 and 519 unrelated individuals respectively. The response variable is the log-transformed vitamin D level.

TABLE 2
*Prediction Errors for the Framingham Dataset*

| Method | Prediction Error | No. of Identified SNPs |
|---|---|---|
| LASSO | .2560 | 14 |
| $\text{LASSO}_a$ | .2085 | 5 |
| $\text{LASSO}_s$ | .2066 | 25 |
| EBVS | .2078 | 1 |
| $\text{EBVS}_i$ | .2121 | 5 |

TABLE 3
*Results of Analyzing the Framingham Data Using LASSO, Adaptive LASSO, Scaled LASSO, EBVS, and EBVS$_i$*

| | | Chromosome-SNP | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1-3887 | 4-0894 | 4-1174 | 5-2773 | 8-5143 | 17-3907 | 17-9089 |
| $\hat{\beta}$ | LASSO | .0412 | 0 | .0355 | .0402 | 0 | 0 | 0 |
| | $\text{LASSO}_a$ | .1521 | 0 | .0434 | .1539 | -.0200 | 0 | .0167 |
| | $\text{LASSO}_s$ | .0990 | -.0112 | .0528 | .1366 | -.0207 | 0 | .0294 |
| | EBVS | 0 | 0 | 0 | .3778 | 0 | 0 | 0 |
| | $\text{EBVS}_i$ | .2417 | -.0542 | 0 | .3047 | -.0857 | .1093 | 0 |
| $p$-value* | LASSO | .2694 | 1 | 1 | .6050 | 1 | 1 | 1 |
| | $\text{LASSO}_a$ | .2060 | 1 | 1 | .0031 | 1 | 1 | 1 |
| | $\text{LASSO}_s$ | 1 | 1 | 1 | .0328 | 1 | 1 | 1 |
| | EBVS | .3138 | 1 | 1 | .0187 | 1 | 1 | 1 |
| | $\text{EBVS}_i$ | .0837 | 1 | 1 | .0034 | 1 | 1 | 1 |
| $\zeta$ | EBVS | .1277 | .0133 | .0347 | .9976 | .0981 | .0869 | .0966 |
| | $\text{EBVS}_i$ | .7609 | .5275 | .3269 | .9718 | .7464 | .8450 | .0009 |

* p-values were calculated using the multi-sample-split method.

We applied lasso, adaptive lasso, scaled lasso, EBVS, and $\text{EBVS}_i$ to the training dataset, and calculated the prediction errors using the test dataset. The results are reported in Table 2. While identifying much more SNPs than all other methods, lasso reported the largest prediction error. Other than scaled lasso ($\text{LASSO}_s$), EBVS has the smallest prediction error though it identified only one SNP. Adaptive lasso ($\text{LASSO}_a$) and $\text{EBVS}_i$ each identified five SNPs, and their prediction errors are slightly higher than that of EBVS.

Presented in Table 3 are the seven SNPs identified to have non-zero regression coefficients by adaptive lasso, EBVS, and $\text{EBVS}_i$. Each SNP is identified by the chromosome it resides in and four digits. The only SNP, 5-2773, which was identified by EBVS, was identified by all other methods. While adaptive lasso and $\text{EBVS}_i$ each identified five SNPs with non-zero regression coefficients, there are only three commonly identified SNPs, i.e., 1-3887, 5-2773, and 8-5143. The two SNPs on chromosome 17, i.e., 17-3907 identified by $\text{EBVS}_i$ and SNP 17-9089 identified by EBVS, neighbor each other with $16k$ bases in between. However the two SNPs on chromosome 4 are far apart from each other.

As in the previous section, we also took the multi-sample-split method to calculate $p$-values based on 50 sample splits for all methods. When we followed Benjamini and Hochberg (1995) to control FDR at 0.1, none of these methods reported any significant SNPs, though adaptive lasso and $\text{EBVS}_i$ reported SNP

5-2773 with the $p$-value as small as 0.0031 and 0.0034 respectively. Instead, when controlling $\widehat{FDR}(\kappa) \leq 0.1$ for both EBVS and EBVS$_i$, EBVS identified only SNP 5-2773, and EBVS$_i$ identified both SNP 5-2773 and 17-3907, with $\kappa = 0.8$. Note that SNP 17-3907 is one of the neighboring pair on chromosome 17. As shown in the simulation studies, $\widehat{FDR}(\kappa)$ usually overestimated $FDR(\kappa)$, so we expect that $FDR(.08) < 0.1$ for both EBVS and EBVS$_i$.

## 6. Discussion

We intend to extend empirical Bayes thresholding (Johnstone and Silverman, 2004) for high-dimensional dependent data, allowing incorporation of complicated *a priori* information on model parameters. An iterative conditional modes/medians (ICM/M) algorithm is proposed to cycle through each coordinate of the parameters for a Bayesian update conditional on all other coordinates. The idea of cycling through coordinates has been revived recently for analyzing high dimensional data. For example, the coordinate descent algorithm has been suggested to obtain penalized least squares estimates, see Fu (1998), Daubechies et al. (2004), Wu and Lange (2008), and Breheny and Huang (2011). However, direct application of the coordinate descent algorithm here is challenged with the spike-and-slab posteriors.

Without *a priori* information other than that regression coefficients are sparse, many lasso-type methods have been proposed with some tuning parameters. It is difficult to select a value for the tuning parameters, and in practice the cross-validation method is widely used. However, high-dimensional data are usually of small sample sizes, and available model fitting algorithms demand intensive computation, both of which disfavor the cross-validation method. In particular, when genome-wide association studies focus more and more on complex diseases associated with rare variants (Nawy, 2012), the limited data usually contain large number of SNPs which differ in a small number of individuals. It is almost infeasible to take a cross-validation method as the small number of unique individuals for a rare variant is more likely to be included in the same fold. Instead, the proposed ICM/M algorithm obtains data-driven hyperparameters via conditional modes, which takes full advantage of each observation in the small sample.

With a large number of predictors and complicated correlation between estimates, classical $p$-values are difficult to compute and it is therefore challenging to evaluate the significance of selected predictors. Wasserman and Roeder (2009), and Meinshausen et al. (2009) recently proposed to calculate $p$-values by splitting the samples. That is, when a sample is split into two folds, one fold is used as the training data to select variables, and the other is used to calculate $p$-values of selected variables. Similar to applying the cross-validation method, splitting samples significantly reduces the power of variable selection and $p$-value calculation, especially for high-dimensional data of small sample sizes. Again, it is almost not feasible to apply such a splitting method to genome-wide association studies with rare variants.

As shown in Section 4, an Ising model as (4.2) can be used to model *a priori* graphical information on predictors. Maximizing pseudo-likelihood approach is utilized to obtain the conditional mode of the Ising model parameters, and therefore the ICM/M algorithm can be easily implemented. Indeed, at each iteration of the ICM/M algorithm, we cycle through all parameters by obtaining conditional modes/medians of one parameter (or a set of parameters), and therefore, many classical approximation methods for low-dimensional issues may be used to simplify the implementation. On the other hand, the Ising prior (4.2) can also be modified to incorporate more complicated *a priori* information on predictors. For example, we may multiply a weight $w_{ij}$ to the interaction $\tau_i\tau_j$ to model the known relationship between the $i$-th and $j$-th predictors. A copula model may be established to model more complicated graphical relationship between the predictors.

For high-dimensional data, stochastic search has been employed to implement Bayesian variable selection, see Hans et al. (2007), Bottolo and Richardson (2010), Li and Zhang (2010), Stingo et al. (2011), and others. The reviewers pointed out that Rockova and George (2014) recently proposed EMVS as an EM approach for rapid Bayesian variable selection. EMVS assumes the "spike-and-slab" Gaussian mixture prior on each $\beta_j$,

$$\beta_j|\omega_j \sim (1-\omega_j)N(0,\nu_0\sigma^2) + \omega_j N(0,\nu_1\sigma^2),$$

where $\omega_j$ is a prior probability, $\nu_1$ takes either a prespecified large value or a $g$-prior, and $\nu_0$ is suggested to explore a sequence of positive values with $\nu_0 < \nu_1$. With an absolutely continuous spike, EMVS estimates $\omega_j$ at the E-step, and estimates $\beta_j$ at the M-step. Note that a positive $\nu_0$ will not automatically yield a sparse estimate of $\beta$, which has to be sparsified using a prespecified threshold. However, the ICM/M algorithm estimates a common $\omega$ based on a conditional mode, and estimates $\beta_j$ based on a conditional median which enables variable selection following Johnstone and Silverman (2004). We also propose a local posterior probability to evaluate the importance of the predictor, which helps control the false discovery rate.

## Acknowledgements

This work was partially supported by NSF CAREER award IIS-0844945, U01CA128535 from the National Cancer Institute, and the Cancer Care Engineering project at the Oncological Science Center of Purdue University. We would like to thank the Editor and the Associate Editor for their insightful comments on the paper, which led to improvement of the manuscript.

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64178. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University.

## Appendix A: Technical details of the ICM/M algorithms

### *A.1. The algorithm in Section 3.1*

Given $\hat{\beta}^{(k)}$, $\hat{\sigma}^{(k)}$, and $\hat{\omega}^{(k)}$ from the $k$-th iteration, the $(k+1)$-st iteration of ICM/M algorithm can proceed in the order of $\hat{\beta}_1^{(k+1)}, \ldots, \hat{\beta}_p^{(k+1)}$, $\hat{\sigma}^{(k+1)}$, and $\hat{\omega}^{(k+1)}$, based on their fully conditional distributions.

Let

$$\begin{cases} \tilde{\mathbf{Y}}_j = \mathbf{Y} - \sum_{l=1}^{j-1} \mathbf{X}_l \beta_l^{(k+1)} - \sum_{l=j+1}^{p} \mathbf{X}_l \beta_l^{(k)}, \\ z_j = \mathbf{X}_j^t \tilde{\mathbf{Y}}_j / (\hat{\sigma}^{(k)} \sqrt{n-1}). \end{cases}$$

Following Proposition 3.1, $\hat{\beta}_j^{(k+1)}$ is updated as the median value of its posterior distribution conditional on $(z_j, \hat{\omega}^{(k)}, \hat{\sigma}^{(k)})$.

Let

$$\begin{aligned} \tilde{F}^{(k+1)}(0|z_j) &= P(\beta_j \geq 0 | z_j, \hat{\omega}^{(k)}, \hat{\sigma}^{(k)}) \\ &= \frac{1 - \Phi(0.5 - z_j)}{[1 - \Phi(z_j + 0.5)]e^{z_j} + \Phi(z_j - 0.5)}, \end{aligned}$$

and $\omega_j = P(\beta_j \neq 0 | z_j, \hat{\omega}^{(k)}, \hat{\sigma}^{(k)})$ which can be calculated as follows,

$$\omega_j^{-1} = 1 + 4(1/\hat{\omega}^{(k)} - 1) \left( \frac{\Phi(z_j - 0.5)}{\phi(z_j - 0.5)} + \frac{1 - \Phi(z_j + 0.5)}{\phi(z_j + 0.5)} \right)^{-1}.$$

If $z_j > 0$, as shown in Johnstone and Silverman (2005), the posterior median $\hat{\beta}_j^{(k+1)}$ is zero if $\omega_j \tilde{F}^{(k+1)}(0|z_j) \leq 0.5$; otherwise,

$$\hat{\beta}_j^{(k+1)} = \frac{\hat{\sigma}^{(k)}}{\sqrt{n-1}} \left\{ z_j - 0.5 - \Phi^{-1} \left( \frac{[1 - \Phi(z_j + 0.5)]e^{z_j} + \Phi(z_j - 0.5)}{2\omega_j} \right) \right\}.$$

If $z_j < 0$, $\hat{\beta}_j^{(k+1)}$ can be calculated on the basis of its antisymmetry property. That is, when a function $\hat{\beta}(z_j) = \hat{\beta}^{(k+1)}$ is defined, then $\hat{\beta}(-z_j) = -\hat{\beta}(z_j)$.

The conditional mode $\hat{\sigma}^{(k+1)}$ can be easily derived following the fact that $\hat{\sigma}^{(k+1)} = \text{mode}(\sigma | \mathbf{Y}, \mathbf{X}, \hat{\beta}^{(k+1)})$, and the conditional mode $\hat{\omega}^{(k+1)}$ can be easily derived following the fact that $\hat{\omega}^{(k+1)} = \text{mode}(\omega | \hat{\beta}^{(k+1)})$.

### *A.2. The algorithm in Section 4.1*

Following Proposition 4.1, $\hat{\beta}_j^{(k+1)}$ is updated as the median value of its posterior distribution conditional on $(z_j, \hat{\varpi}_j, \hat{\sigma}^{(k)})$, where $\hat{\varpi}_j$ is calculated as follows,

$$\hat{\varpi}_j^{-1} = 1 + \exp\left\{ -\hat{a}^{(k+1)} - \hat{b}^{(k+1)} \sum_{l:<j,k>\in E} \hat{\tau}_l \right\},$$

with $\hat{\tau}_l = I\{\hat{\beta}_l^{(k+1)} \neq 0\}$ for $l = 1, \ldots, j-1$; and $\hat{\tau}_l = I\{\hat{\beta}_l^{(k)} \neq 0\}$ for $l = j+1, \ldots, p$.

The conditional median $\hat{\beta}_j^{(k+1)}$ can be calculated following A.1, except that the posterior probability $\omega_j = P(\beta_j \neq 0 | z_j, \hat{\varpi}_j, \hat{\sigma}^{(k)})$ should be updated as follows,

$$\omega_j^{-1} = 1 + 4(1/\hat{\varpi}_j - 1)\left(\frac{\Phi(z_j - 0.5)}{\phi(z_j - 0.5)} + \frac{1 - \Phi(z_j + 0.5)}{\phi(z_j + 0.5)}\right)^{-1}.$$

## References


Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32:870–897. MR2065192

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300. MR1325392

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society Series D (The Statistician)*, 24:179–195.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B*, 48:259–302. MR0876840

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5:583–618. MR2719668

Breheny, P. and Huang, J. (2011). Cooridinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5:232–253. MR2810396

Brenner, V., Lindauera, K., Parkara, A., Fordhama, J., Hayesa, I., Stowa, M., Gamaa, R., Pollocka, K., and Jupp, R. (2001). Analysis of cellular adhesion by microarray expression profiling. *Journal of Immunological Methods*, 250:15–28.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, 57:473–484.

Cupples, L. A., Arruda, H. T., Benjamin, E. J., and *et al.* (2007). The framingham heart study 100k snp genome-wide association study resource: Overview of 17 phenotype working group reports. *BMC Medical Genetics*, 8(Suppl 1):S1.

Daubechies, I., Defrise, M., and Mol, C. D. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commications on Pure and Applied Mathematics*, 57:1413–1457. MR2077704

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455. MR1311089

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103. MR1997066

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360. MR1946581

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416. MR1646710

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of American Statistical Association*, 85:398–409.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163.

Geyer, C. J. and Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series B*, 54:657–699. MR1185217

Guyon, X. and Kunsch, H. R. (1992). Asymptotic comparison of estimators in the ising model. In Barone, P., Frigessi, A., and Piccioni, M., editors, *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, pages 177–198. Springer, New York. MR1188486

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102:507–516. MR2370849

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33:730–773. MR2163158

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of Landon Series A*, 196:453–461. MR0017504

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequence. *The Annals of Statistics*, 32:1594–1649. MR2089135

Johnstone, I. M. and Silverman, B. W. (2005). Ebayesthresh: R programs for empirical bayes thresholding. *Journal of Statistical Software*, 12:1–38.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The Annals of Applied Statistics*, 4:1498–1516. MR2758338

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with application in genomics. *Journal of the American Statistical Association*, 105:1202–1214. MR2752615

Liang, G. and Yu, B. (2003). Maximum pseudo likelihood estimation in network tomography. *IEEE Transactions on Signal Processing*, 51:2043–2053.

Mase, S. (2000). Marked gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Mathematische Nachrichten*, 209:151–169. MR1734363

Meinshausen, N., Meier, L., and Buehlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681. MR2750584

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1036. MR0997578

Nawy, T. (2012). Rare variants and the power of association. *Nature Methods*, 9:324.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176.

Onsager, L. (1943). Crystal statistics. i. A two-dimensional model with an order-disorder transition. *Physical Review*, 65:117–149. MR0010315

Pan, W., Xie, B., and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66:474–484. MR2758827

Rockova, V. and George, E. I. (2014). Incorporating predictor network in penalized regression with application to microarray data. *Journal of the American Statistical Association*, 109:828–846. MR3223753

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5:1978–2002. MR2884929

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99:879–898. MR2999166

Syed, V. and Hecht, N. B. (1997). Up-regulation and down-regulation of genes expressed in cocultures of rat sertoli cells and germ cells. *Molecular Reproduction and Development*, 47:380–389.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B*, 58:267–288. MR1379242

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42. MR2796852

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Annals of Statistics*, 37:2178–2201. MR2543689

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244. MR2415601

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society Series B*, 68:49–67. MR2212574

Zhang, M., Zhang, D., and Wells, M. T. (2010). Generalized thresholding estimators for high-dimensional location parameters. *Statistica Sinica*, 20:911–926. MR2682648

Zhou, X. and Schmidler, S. C. (2009). Bayesian parameter estimation in ising and potts models: A comparative study with applications to protein modeling. Technical report, Duke University.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429. MR2279469