An Iterated Conditional Modes/Medians Algorithm for Empirical Bayes Selection of Massive Variables

Vitara Pungpapong, Min Zhang, Dabao Zhang*

August 22, 2013

Technical Report 13-05

*Vitara Pungpapong (email: vitara@cbs.chula.ac.th) is Lecturer, Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand. Min Zhang (minzhang@stat.purdue.edu) and Dabao Zhang (zhangdb@stat.purdue.edu) are Associate Professors, Department of Statistics, Purdue University, West Lafayette, IN 47907, USA. This work was partially supported by NSF CAREER award IIS-0844945 and the Cancer Care Engineering project at the Oncological Science Center of Purdue University.

Abstract

Empirical Bayes methods are privileged in data mining because they can absorb prior information on model parameters and are free of choosing tuning parameters. We proposed 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 algorithm is constructed on the basis of iteratively minimizing a conditional loss function. 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 very competitive performance of the proposed method, even in the simple case of selecting massive regression predictors.

Key Words: High Dimensional Data; Prior; Sparsity; Structured Variables

1 Introduction

Selecting variables out of massive candidates 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 pierce 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 [11]. Many methods have been developed to take full advantage of this sparsity assumption, mostly built upon thresholding procedures [9], see Tibshirani [31], Fan and Li [11], and others.

Recently many efforts have been devoted to selecting variables from massive candidates by incorporating rich *a priori* information accumulated from historical researches or practices. For example, Yuan and Lin [35] defined group-wise norms for grouped variables. For graph-structured variables, Li and Li [20] and Pan et al. [29] proposed to use Laplacian matrices and L_γ norms, respectively. Li and Zhang [21] and Stingo et al. [30] 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, for example, George and McCulloch [13], Carlin and Chib [6], Li and Zhang [21], and Stingo et al. [30]. However, MCMC algorithms are computationally intensive and may be difficult to obtain appropriate hyperparameters. On the other hand, penalty-based variable selection usually demands predetermination of certain tuning parameters [e.g. 31, 11, 35, 20, 29], 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.

Empirical Bayes methods are privileged 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 [19] modeled the sparse normal means using a spike and slab prior. The mixing rate of the 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 [19], 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 [4], 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 [19], 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 [21], we illustrate such strength of ICM/M.

The rest of this paper is organized as follows. In the next section, we will describe the general idea behind the empirical Bayes variable selection (EBVS), and propose the ICM/M algorithm for 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, taking the only assumption that non-zero regression coefficients are few. Shown in Section 4 is the ICM/M algorithm when incorporating *a priori* information on graphical relationship between the predictors. Simulation studies are carried out in both Section 3 and 4 to evaluate the performance of the corresponding ICM/M algorithms. An application to a real dataset in genome-wide association study is presented in Section 5. We conclude this paper with a discussion in Section 6.

2 The General Idea

2.1 Empirical Bayes Variable Selection

Consider a general variable selection issue presented with a likelihood function,

$$\mathcal{L}(\mathbf{Y}; \mathbf{X}\beta; \phi), \tag{1}$$

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

A typical variable selection task is to identify non-zero components in β, that is, to select variables, out of the p candidates, with effects on $\mathbf{Y}$. 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 $\tau = (\tau_1, \tau_2, \cdots, \tau_p)^t$.

Here we consider an empirical Bayes variable selection, which assumes priors,

$$\begin{cases} \beta \sim \pi(\beta|\tau, \psi_1) \times \pi(\tau|\psi_2), \\ \phi \sim \pi(\phi|\psi_3), \end{cases} \tag{2}$$

where ψ_1, ψ_2, and ψ_3 are hyperparameters. Let $\psi = (\psi_1^t, \psi_2^t, \psi_3^t)^t$, then a maximum a posteriori (MAP) estimate is

$$\hat{\psi} = \arg\max_{\psi} \int\int \mathcal{L}(\mathbf{Y}; \mathbf{X}\beta; \phi) \left[\sum_{\tau} \pi(\beta|\tau, \psi_1)\pi(\tau|\psi_2) \right] \pi(\phi|\psi_3) d\beta d\phi. \tag{3}$$

An empirical Bayes variable selection can proceed as finding an estimate $\hat{\beta} = \hat{\beta}(\mathbf{Y}, \mathbf{X}, \hat{\psi})$, together with $\hat{\phi} = \hat{\phi}(\mathbf{Y}, \mathbf{X}, \hat{\psi})$, such that,

$$(\hat{\beta}, \hat{\phi}) = \arg\min_{\tilde{\beta}, \tilde{\phi}} \left\{ E\left[E\left[L(\tilde{\beta}(\mathbf{Y}, \mathbf{X}, \hat{\psi}), \tilde{\phi}(\mathbf{Y}, \mathbf{X}, \hat{\psi}); \beta, \phi) \middle| \beta, \phi, \hat{\psi} \right] \middle| \hat{\psi} \right] \right\}, \tag{4}$$

where L is a loss function and can be set up as follows, with $\tilde{\beta} = \tilde{\beta}(\mathbf{Y}, \mathbf{X}, \hat{\psi})$ and $\tilde{\phi} = \tilde{\phi}(\mathbf{Y}, \mathbf{X}, \hat{\psi})$,

$$L(\tilde{\beta}, \tilde{\phi}; \beta, \phi) = \|\tilde{\beta}(\mathbf{Y}, \mathbf{X}, \hat{\psi}) - \beta\|_1 + \|\tilde{\phi}(\mathbf{Y}, \mathbf{X}, \hat{\psi}) - \phi\|_0, \tag{5}$$

where $\|\cdot\|_1$ refers to the l_1 norm, $\|\cdot\|_0$ refers to the l_0 norm. Here the zero-one loss on ϕ follows the iterated conditional modes by Besag [4] to analyze massive data. As shown by

Johnstone and Silverman [19], the absolute-error loss on β results in a thresholding estimate of high-dimensional β, which is adaptive to signal sparsity when constructed with appropriate priors on β.

The Bayesian risk in (4) can be rewritten as

$$\begin{aligned} R(\tilde{\beta},\tilde{\phi}|\hat{\psi}) &= E\left[\sum_j E\left[|\tilde{\beta}_j(\mathbf{Y},\mathbf{X},\hat{\psi})-\beta_j|\big|\mathbf{Y},\mathbf{X},\hat{\psi}\right]\right. \\ &\quad \left.+\sum_j E\left[I\{\tilde{\phi}_j(\mathbf{Y},\mathbf{X},\hat{\psi})\neq\phi_j\}\big|\mathbf{Y},\mathbf{X},\hat{\psi}\right]\Big|\hat{\psi}\right], \end{aligned} \tag{6}$$

where β_j and ϕ_j refer to the j-th components of β and ϕ respectively. When considering the inner expectation, we observe that

$$\hat{\beta}_j=\hat{\beta}_j(\mathbf{Y},\mathbf{X},\hat{\psi})=\text{median}(\beta_j|\mathbf{Y},\mathbf{X},\hat{\psi}) \tag{7}$$

minimizes the first part, and

$$\hat{\phi}_j=\hat{\phi}_j(\mathbf{Y},\mathbf{X},\hat{\psi})=\text{mode}(\phi_j|\mathbf{Y},\mathbf{X},\hat{\psi}) \tag{8}$$

minimizes the second part when the corresponding posterior is unimodal.

2.2 Iterated Conditional Modes/Medians

We here consider the empirical Bayes variable selection by minimizing a Bayes risk with the loss function defined as

$$L_F(\tilde{\beta},\tilde{\phi},\tilde{\psi};\beta,\phi,\psi)=\|\tilde{\beta}-\beta\|_1+\|\tilde{\phi}-\phi\|_0+\|\tilde{\psi}-\psi\|_0. \tag{9}$$

Indeed, minimizing the corresponding Bayes risk is subject to finding $\tilde{\beta}=\tilde{\beta}(\mathbf{Y},\mathbf{X})$, $\tilde{\phi}=\tilde{\phi}(\mathbf{Y},\mathbf{X})$, and $\tilde{\psi}=\tilde{\psi}(\mathbf{Y},\mathbf{X})$ which minimize

$$E\left[L_F(\tilde{\beta},\tilde{\phi},\tilde{\psi};\beta,\phi,\psi)\big|\mathbf{Y},\mathbf{X}\right]. \tag{10}$$

However, for even moderately complicated model (1), minimizing (10) for optimal $(\tilde{\beta},\tilde{\phi},\tilde{\psi})$ can be difficult as it involves high-dimensional integration.

Define a conditional loss function,

$$\begin{aligned} L_C(\tilde{\beta},\tilde{\phi},\tilde{\psi};\beta,\phi,\psi) &= \sum_j E\left[|\tilde{\beta}_j-\beta_j|\big|\mathbf{Y},\mathbf{X},\beta_{-j},\phi,\psi\right] \\ &\quad+\sum_j E\left[I\{\tilde{\phi}_j\neq\phi_j\}\big|\mathbf{Y},\mathbf{X},\beta,\phi_{-j},\psi\right] \\ &\quad+\sum_j E\left[I\{\tilde{\psi}_j\neq\psi_j\}\big|\mathbf{Y},\mathbf{X},\beta,\phi,\psi_{-j}\right], \end{aligned} \tag{11}$$

where $\tilde{\beta}_j$, $\tilde{\phi}_j$, and $\tilde{\psi}_j$ are the j-th components of $\tilde{\beta}$, $\tilde{\phi}$, and $\tilde{\psi}$ respectively; β_{-j} refers to β excluding the j-th component, ϕ_{-j} refers to ϕ excluding the j-th component, and ψ_{-j} refers to ψ excluding the j-th component. Then,

$$E\left[L_F(\tilde{\beta},\tilde{\phi},\tilde{\psi};\beta,\phi,\psi)\big|\mathbf{Y},\mathbf{X}\right]=E\left[L_C(\tilde{\beta},\tilde{\phi},\tilde{\psi};\beta,\phi,\psi)\big|\mathbf{Y},\mathbf{X}\right]. \tag{12}$$

We here consider iteratively minimizing (11) with an initial point at $\hat{\beta}^{(0)}=\hat{\beta}^{(0)}(\mathbf{Y},\mathbf{X})$, $\hat{\phi}^{(0)}=\hat{\phi}^{(0)}(\mathbf{Y},\mathbf{X})$, and $\hat{\psi}^{(0)}=\hat{\psi}^{(0)}(\mathbf{Y},\mathbf{X})$. That is, given a point $(\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)})$, we minimize the conditional loss function $L_C(\tilde{\beta},\tilde{\phi},\tilde{\psi};\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)})$ for an optimal point $(\hat{\beta}^{(k+1)},\hat{\phi}^{(k+1)},\hat{\psi}^{(k+1)})$, which suggests

$$\begin{cases} \hat{\beta}_j^{(k+1)}=\hat{\beta}_j(\hat{\beta}_{-j}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)})=\text{median}(\beta_j|\mathbf{Y},\mathbf{X},\hat{\beta}_{-j}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)}), \\ \hat{\phi}_j^{(k+1)}=\hat{\phi}_j(\hat{\beta}^{(k)},\hat{\phi}_{-j}^{(k)},\hat{\psi}^{(k)})=\text{mode}(\phi_j|\mathbf{Y},\mathbf{X},\hat{\beta}^{(k)},\hat{\phi}_{-j}^{(k)},\hat{\psi}^{(k)}), \\ \hat{\psi}_j^{(k+1)}=\hat{\psi}_j(\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}_{-j}^{(k)})=\text{mode}(\psi_j|\mathbf{Y},\mathbf{X},\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}_{-j}^{(k)}). \end{cases} \tag{13}$$

When the sequence of $\{(\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)}):k=1,2,\cdots\}$ converges to $(\hat{\beta},\hat{\phi},\hat{\psi})$, then $\hat{\beta}=\hat{\beta}(\mathbf{Y},\mathbf{X})$, $\hat{\phi}=\hat{\phi}(\mathbf{Y},\mathbf{X})$, and $\hat{\psi}=\hat{\psi}(\mathbf{Y},\mathbf{X})$.

The above iterative method suggests an conditional median for β_j, and conditional modes for ϕ_j and ψ_j respectively, hereafter named the iterated conditional medians/modes (ICM/M) algorithm for implementing the empirical Bayes variable selection. Indeed, each component of $(\hat{\beta}^{(k+1)},\hat{\phi}^{(k+1)},\hat{\psi}^{(k+1)})$ is a Bayesian update of the corresponding component of $(\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)})$ conditional on all other components. Obviously, a consistent initial point $(\hat{\beta}^{(0)},\hat{\phi}^{(0)},\hat{\psi}^{(0)})$ leads to a well-established update $(\hat{\beta},\hat{\phi},\hat{\psi})$.

Note that the iterative method in (13) is well suited for parallel computing in the case of high-dimensional data. In the rest of the paper, we will focus on ICM/M algorithm for non-parallel computing. To accelerate update of the sequence $\{(\hat{\beta}^{(k)},\hat{\phi}^{(k)},\hat{\psi}^{(k)}):k=1,2,\cdots\}$, we

can sequentially update each component of $(\hat{\beta}^{(k+1)}, \hat{\phi}^{(k+1)}, \hat{\psi}^{(k+1)})$ conditional on the most recent values of all other components. Specifically, when $\hat{\beta}^{(k)}$, $\hat{\phi}^{(k)}$ and $\hat{\psi}^{(k)}$ have been obtained in the k-th iteration, the $(k+1)$-st iteration proceeds as follows,

$$\begin{cases} \hat{\beta}_j^{(k+1)} = \text{median}(\beta_j | \mathbf{Y}, \mathbf{X}, \hat{\beta}_{1:(j-1)}^{(k+1)}, \hat{\beta}_{(j+1):p}^{(k)}, \hat{\phi}^{(k)}, \hat{\psi}^{(k)}), \\ \hat{\phi}_j^{(k+1)} = \text{mode}(\phi_j | \mathbf{Y}, \mathbf{X}, \hat{\beta}^{(k+1)}, \hat{\phi}_{1:(j-1)}^{(k+1)}, \hat{\phi}_{(j+1):q}^{(k)}, \hat{\psi}^{(k)}), \\ \hat{\psi}_j^{(k+1)} = \text{mode}(\psi_j | \mathbf{Y}, \mathbf{X}, \hat{\beta}^{(k+1)}, \hat{\phi}^{(k+1)}, \hat{\psi}_{1:(j-1)}^{(k+1)}, \hat{\psi}_{(j+1):r}^{(k)}). \end{cases} \tag{14}$$

When each $\hat{\beta}_j^{(k+1)}$ is also obtained as an conditional mode, the above algorithm concurs with the iterated conditional modes(ICM) algorithm by Besag [4]. However, calculation of conditional mode for $\hat{\beta}_j^{(k+1)}$ is either infeasible or practically undesirable (due to lack of variable selection function). Indeed, Bayesian or empirical Bayes variable selection usually follows a spike and slab prior on each β_j [e.g. 25, 17], and it induces a spike and slab posterior for each β_j. While it is infeasible to obtain the mode of such a spike and slab posterior, its median can be zero and therefore allows to select the median probability model as suggested by Barbieri and Berger [1]. As shown in later sections, ICM/M algorithm allows an easy extension of the (generalized) empirical Bayes thresholding methods by Johnstone and Silverman [19] to dependent data.

2.3 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, our objective of high-dimensional data analysis usually 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 screening out a short list of candidates for follow-up studies.

For Bayesian approach, inference to 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 in the case that predictors are highly correlated (which usually occurs in large p small n data set. For example, suppose

predictor X_1 and X_2 are linearly dependent and both predictors are associated to 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,

$$\zeta_j = P(\beta_j \neq 0 | \mathbf{Y}, \mathbf{X}, \hat{\beta}_{-j}, \hat{\phi}, \hat{\psi}). \tag{15}$$

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 would be how large the list of important predictors should be. In many literatures, the numbers of important variables reported are arbitrary. For instance, some laboratory might be interested in looking at, say, the top ten genes. Typically, however, there is an interest to create the list such that errors are controlled in some way such as type-I and type-II errors [10]. False discovery rate (FDR) control is widely used in high-dimensional data since it is less conservative and has more power than controlling familywise error rate [2].

With the local posterior probability ζ and assumption that true β is known, we can report a list containing predictors having the posterior probability greater than some bound κ, $0 \leq \kappa < 1$. Given the data, true FDR can be computed as

$$FDR(\kappa) = \sum_{j=1}^{p} I\{\beta_j = 0, \zeta_j > \kappa\} \Big/ \sum_{j=1}^{p} I\{\zeta_j > \kappa\}. \tag{16}$$

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

$$\widehat{FDR}(\kappa) = \sum_{j=1}^{p} (1 - \zeta_j) I\{\zeta_j > \kappa\} \Big/ \sum_{j=1}^{p} I\{\zeta_j > \kappa\}. \tag{17}$$

We then can select predictors to report by controlling $\widehat{FDR}(\kappa)$ at a desired level. $\widehat{FDR}(\kappa)$ is just an approximation because it depends on the accuracy of the fitted model. Careful modeling and diagnostic checking can reduce the effect of this approximation [27].

3 Selection of Sparse Variables

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

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n). \tag{18}$$

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

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

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

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

$$\frac{1}{n-1}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j \sim N\left(\beta_j, \frac{1}{n-1}\sigma^2\right). \tag{19}$$

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

$$\beta_j|\sigma \sim (1-\omega)\delta_0(\beta_j) + \omega\gamma(\beta_j|\sigma), \tag{20}$$

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 β_j is zero with probability $(1-\omega)$ and is drawn from the nonzero part of prior, $\gamma(\cdot|\sigma)$, with probability ω. As suggested by Johnstone and Silverman [19], a heavy-tailed prior such as Laplace distribution can be a good choice for $\gamma(\cdot|\sigma)$, that is,

$$\gamma(\beta_j|\sigma) = \frac{\alpha\sqrt{n-1}}{2\sigma}\exp\left(-\frac{\alpha\sqrt{n-1}}{\sigma}|\beta_j|\right), \tag{21}$$

where $\alpha > 0$ is a scale parameter. We take Jeffreys' prior on σ as $\pi(\sigma) \propto 1/\sigma$ [18].

Note that there is a connection of using Laplace prior and the lasso. Indeed, setting $\omega = 1$ in (20) leads to a lasso estimate with α related to a tuning parameter in the lasso, see details in Tibshirani [31]. Our empirical Bayes variable selection allows a data-driven optimal choice of ω. Indeed, a data-driven optimal α can also be obtained through the conditional mode suggested by (14), 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 [19] 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 (14). Note that $\phi = \sigma$, and $\psi = (\omega, \alpha)$ or $\psi = \omega$ depending on whether α is fixed. Throughout this paper, we fix $\alpha = 0.5$ as suggested by Johnstone and Silverman [19].

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 β_j in (19) and it is therefore easy to calculate $\hat{\beta}_j^{(k+1)}$ as stated below. Indeed, $\hat{\beta}_j^{(k+1)}$ is a (generalized) empirical Bayes thresholding estimator as shown in Johnstone and Silverman [19].

Proposition 3.1. *With pre-specified values of σ and β_{-j}, $\frac{1}{n-1}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j$ is a sufficient statistic for β_j w.r.t the model (18). Furthermore, the iterative conditional median of β_j in the ICM/M algorithm can be constructed as the posterior median of β_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 (18) 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 σ is set to one. The predictors are partitioned into ten blocks, each block including 100 predictors which are

serially correlated at the same level of correlation coefficient ρ. We simulated 100 datasets for each ρ in $\{0, 0.1, 0.2, \cdots, 0.9\}$.

EBVS was compared with two popularly considered approaches, i.e., lasso by Tibshirani [31], and adaptive lasso by Zou [37]. The 10-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 both lasso and adaptive lasso in terms of prediction error rates. In particular, when $\rho \geq 0.3$, EBVS consistently reported median prediction error rates approximately at 1.5. In comparison of lasso and adaptive lasso, adaptive lasso has smaller prediction error rates when $\rho < 0.3$; but lasso has smaller prediction error rates lasso when $\rho > 0.3$.

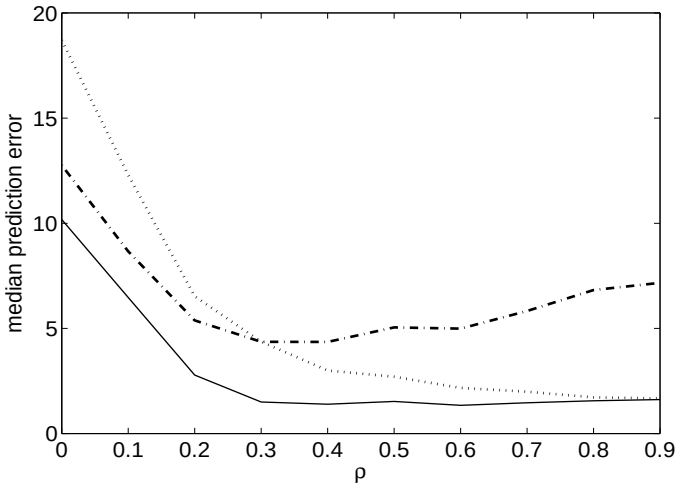

Figure 1: Median prediction errors of lasso (dotted), adaptive lasso (dash-dotted), and EBVS (solid) for simulation study in Section 3.2.

It is known that lasso can inconsistently select variables under certain conditions, and adaptive lasso was proposed for solving this issue [37]. Figure 2 showed that lasso has very high false positive rates (more than 50%), and adaptive lasso significantly lowers 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.

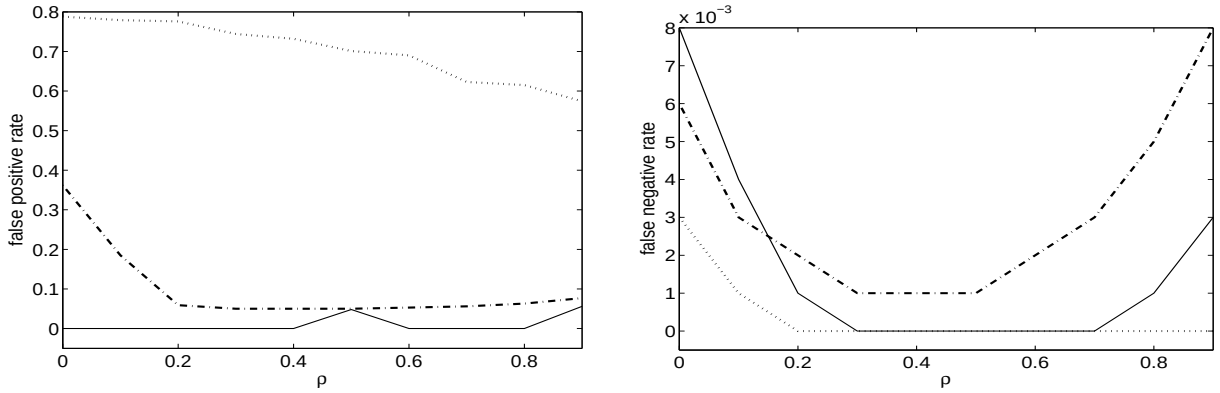

Figure 2: False positive rate (left) and false negative rate (right) of lasso (dotted), adaptive lasso (dash-dotted), and EBVS (solid) for simulation study in Section 3.2.

Recently, Meinshausen et al. [24] 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 ζ_i defined in (15). 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, ζ_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 κ to threshold ζ_i for the purpose of controlling FDR.

4 Selection of Structured Variables

When the information of structural relationship among predictors is available, it is unreasonable to assume independent prior on each $\beta_j, j = 1, ..., p$ as described in previous section. Instead, we introduce an indicator variable $\tau = (\tau_1, ..., \tau_p)^T$ where $\tau_j = I\{\beta_j \neq 0\}$. Then, the prior distribution of $\tilde{\beta}$ is set to be dependent to τ. Specifically, given τ_j, β_j has the mixture

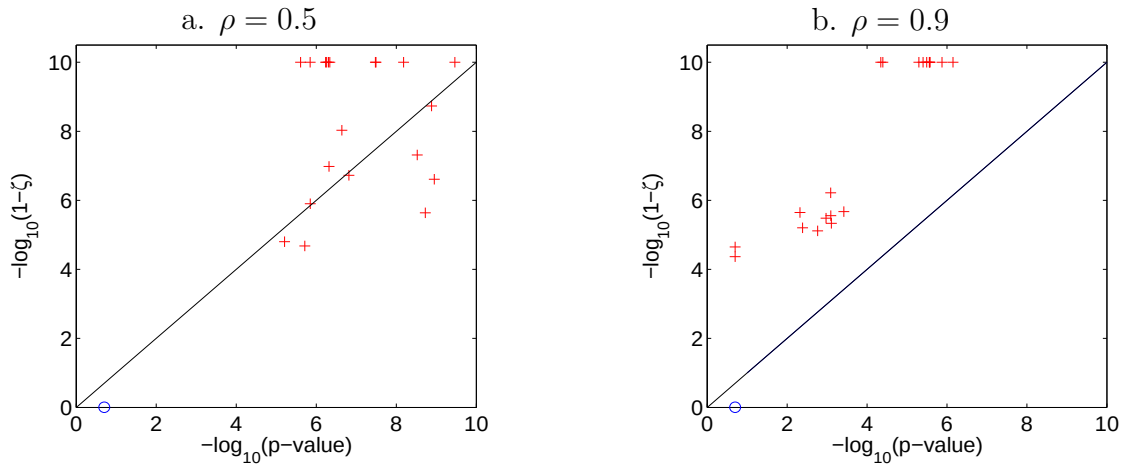

Figure 3: Comparison of the local posterior probabilities (with $-\log_{10}(1-\zeta)$ truncated at 10) and p-values in evaluating variable importance. The results are based on the simulation study of EBVS in Section 3.2. True positives are indicated by crosses and true negatives are indicated by circles.

distribution

$$\beta_j|\tau_j \sim (1-\tau_j)\delta_0(\beta_j) + \tau_j\gamma(\beta_j), \tag{22}$$

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

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 [28] is employed to model the *a priori* information on τ,

$$P(\tau) = \frac{1}{Z(a,b)} \exp\left\{ a\sum_i \tau_i + b \sum_{<i,j>\in E} \tau_i\tau_j \right\}, \tag{23}$$

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 τ_i and τ_j tend to have the same value. When $b < 0$, the interaction is called *antiferromagnetic*, i.e., neighboring τ_i and τ_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 τ_i. That is, if $a+b>0$, τ_i tends to be one; if $a+b<0$, τ_i tends to be zero.

4.1 The Algorithm

Here we will implement 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 for more general priors. With $\alpha = 0.5$, the ICM/M algorithm described in (14) 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 (23). However, it requires to compute $Z(a, b)$ by summing up p-dimensional space of τ, which demands intensive computation especially for a large p. Many methods have been proposed for approximate calculation, see Geyer [14], Geyer and Thompson [15], Zhou and Schmidler [36] and others. Here we will consider the composite likelihood approach [32] 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 [3].

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 [22]. 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 [16, 23, 32]. The implementation of pseudo-likelihood method is fast and straightforward which is feasible for a large scale of graph. 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 $i = 1, \cdots, p$.

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

Proposition 4.1. *With pre-specified values of σ, a, b, and β_{-j}, $\frac{1}{n-1}\mathbf{X}_j^t\tilde{\mathbf{Y}}_j$ is a sufficient statistic for β_j w.r.t the model (18). Furthermore, the iterative conditional median of β_j in the ICM/M algorithm can be constructed as the posterior median of β_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 ϖ_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 (18) with structured predictors, i.e., the values of β_j depend on correlated τ_j. We here consider two different correlation structures of τ_i. Both EBVS and EBVS$_i$ were applied to each simulated dataset. They were compared with two other methods, i.e., lasso and adaptive lasso.

Case I. Markov Chain. For each $j = 1, \cdots, p$, $\beta_j = 0$ if $\tau_j = 0$; and if $\tau_j = 1$, β_j is independently sampled from a uniform distribution on $[0.3, 2]$. The indicator variables τ_1, $\cdots$, τ_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 τ_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 ρ ranging from 0 to 0.9 with step 0.1.

Similar to the simulation study in Section 3.2, the prediction error rates of true parameters are close to the error variance which is one, see Figure 4. EBVS performed slightly better than adaptive lasso, and both performed much better than lasso. Lasso, adaptive lasso, and EBVS all presented varying prediction error rates when ρ goes from 0 to 0.9. However, the prediction error rates of EBVS$_i$ are rather stable for varying values of ρ, and are much smaller than that of the other three methods.

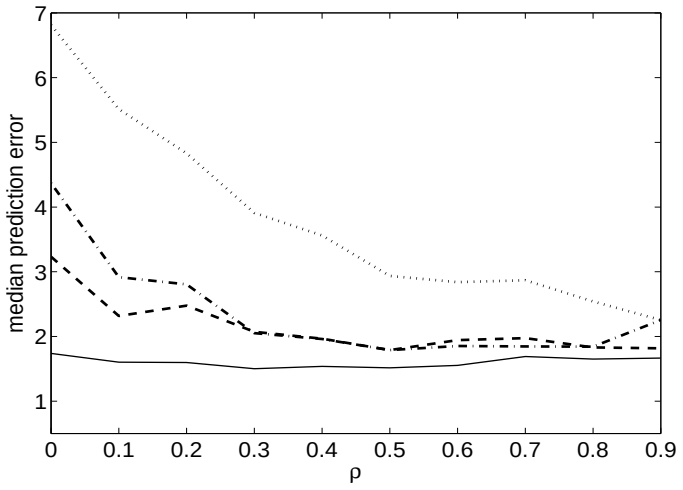

Figure 4: Median prediction errors of lasso (dotted), adaptive lasso (dash-dotted), EBVS (dashed), and EBVS$_i$ (solid) for simulation study of Case I.

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 lowered the false positive rates, which is still more than 10%. Instead both EBVS and EBVS$_i$ reported false positive rates under 10%. Indeed, EBVS reported false positive rates at zero for different values of ρ; and EBVS$_i$ reported false positive rates at zero when $\rho < 0.6$, and and 0.1 when $\rho \geq 0.6$. However, EBVS$_i$ reported false negative rates lower than EBVS. Therefore, EBVS tends to select correct 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, its highly correlated neighboring predictors are preferred to be included in the model as false positives.

Figure 6 shows $FDR(\kappa)$ and $\widehat{FDR}(\kappa)$ of EBVS$_i$ for the models with $\rho = 0.5$ and $\rho = 0.9$

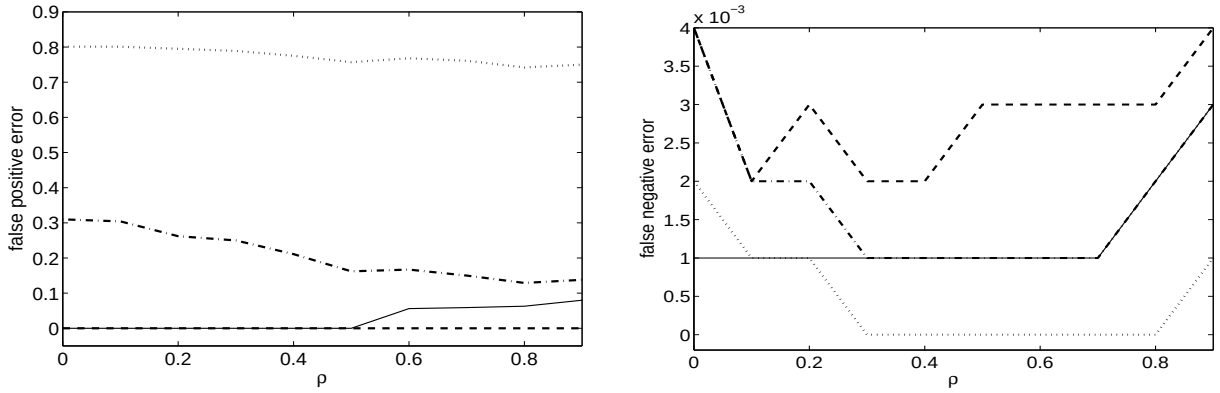

Figure 5: False positive rate (left) and false negative rate (right) of lasso (dotted), adaptive lasso (dash-dotted), EBVS (dashed), and EBVS$_i$ (solid) for simulation study of Case I.

respectively (we also observed that $FDR(\kappa)$ of EBVS is similar to that of EBVS$_i$, which is 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, we would like to list important predictors while controlling FDR at 0.1 for the model with $\rho = 0.9$. We should select κ around 0.1 based on $FDR(\kappa)$. However, we will select κ around 0.4 based on $\widehat{FDR}(\kappa)$, which suggests a true FDR as low as zero.

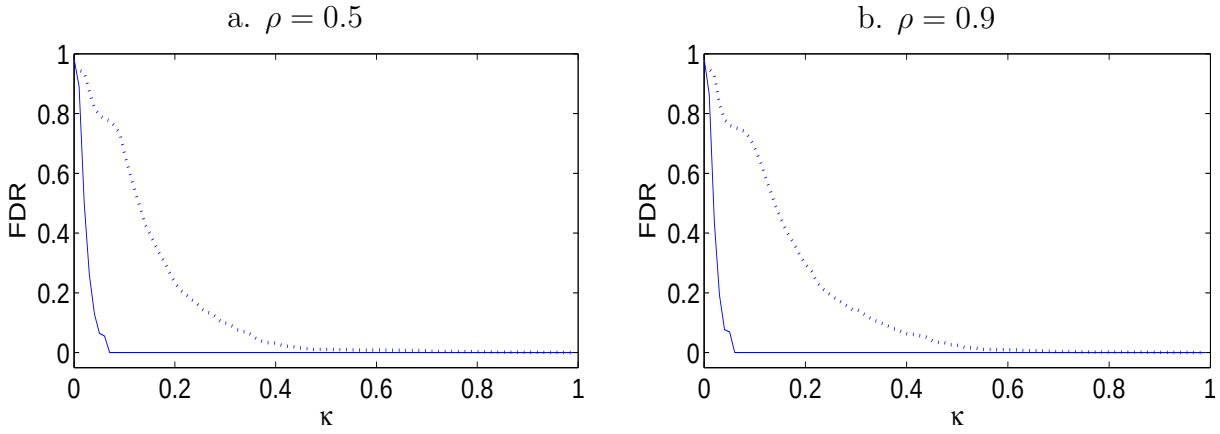

Figure 6: Median true FDR (solid) and estimated FDR (dotted) versus κ plots based on the results from EBVS$_i$ for simulations in Case I.

Plotted in Figure 7 are the p-values, calculated using the multi-sample-split method [24], against ζ_j for each predictor. For both EBVS and EBVS$_i$, ζ_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 ζ for true positives, while both

EBVS and EBVS$_i$ keep true negatives with ζ_j close to zero. Indeed, EBVS produced ζ_j close to 0 for several true positives while EBVS$_i$ produced larger values of ζ_j for these true positives. We then summarize empirically that, by incorporating *a priori* information, EBVS$_i$ has more power to detect true positives than EBVS.

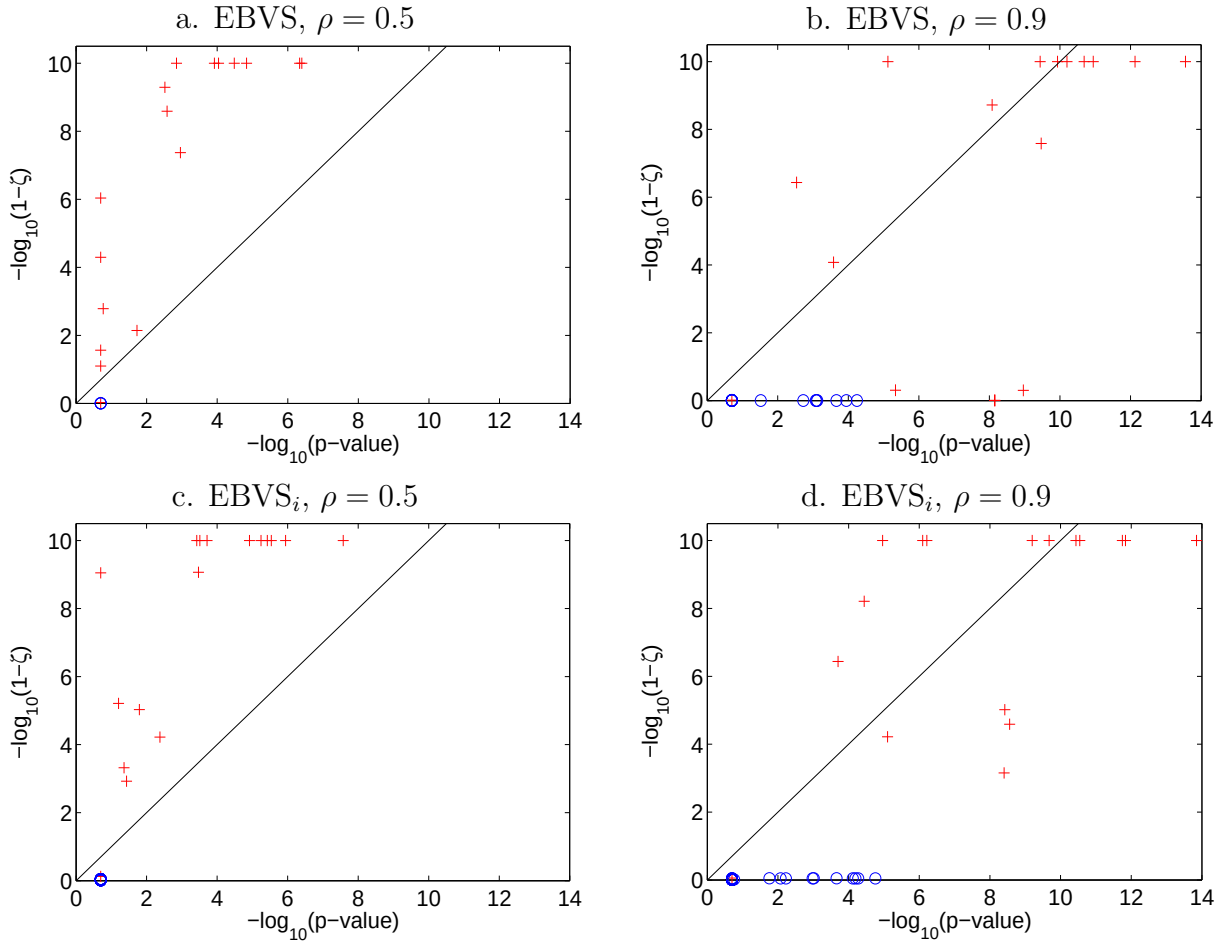

Figure 7: Comparing the plots of local posterior probabilities (with $-\log_{10}(1-\zeta)$ truncated at 10) versus p-values from EBVS and EBVS$_i$ in simulation study of Case I. True positives are indicated by crosses and true negatives are indicated by circles.

Case II. Pathway Information. To mimic a real genome-wide association study (GWAS), we took values of some single nucleotide polymorphisms (SNPs) in Framingham dataset [7] to generate $\mathbf{X}$ in model (18). 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 Framingham dataset were randomly selected out of those SNPs residing in the genetic region.

If no SNPs could be found inside 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 Framingham dataset, and used them to generated predictor values of the training dataset. For the rest of Framingham dataset, we similarly identified 653 unrelated individuals to generate predictor values of the test dataset. Five pathways were assumed to be associated to 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 ranging over $[0.5, 3]$. The error variance is 5. A total of 100 datasets were simulated.

As shown in Table 1, lasso has relatively low prediction error rate. However, its median false positive rate is as high as 69%, much higher than others. Adaptive lasso (LASSO_a), on the other hand, has very large prediction error rate but its false positive rate is much lower 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 reduced the false positive rate from lasso by more than 98%. By incorporating the pathway information using an Ising prior on τ, EBVS_i reported the lowest prediction error rate. Furthermore, EBVS_i compromised between lasso, adaptive lasso, and EBVS to balance well between the false positive rate and false negative rate.

Table 1: Results of Simulation Study on Case II.

Method	Prediction Error (s.e.)	False Positive (s.e.)	False Negative (s.e.)
LASSO	30.6928(.4050)	.6905(.0004)	.0204(.0004)
LASSO_a	206.1994(.5726)	.0744(.0017)	.1266(.0002)
EBVS	95.3686(1.8820)	.0118(.0010)	.0970(.0008)
EBVS_i	21.7731(.2320)	.0308(.0015)	.0394(.0003)

5 Real Data Analysis

Here empirical Bayes variable selection via ICM/M algorithm was applied to publicly available Framingham dataset [7] to find SNPs associated to 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 left, 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 in this analysis is the log-transformed vitamin D level.

We applied lasso, adaptive lasso, EBVS, and EBVS$_i$ to the training dataset, and calculated the prediction error rates 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 rate. EBVS has the smallest prediction error rate though it identified only one SNP. Adaptive lasso (LASSO$_a$) and EBVS$_i$ each identified five SNPs, and their prediction error rates are slightly higher than that of EBVS.

Table 2: Prediction Error Rates for Framingham Dataset.

Method	Prediction Error Rates	No. of Identified SNPs
LASSO	.2560	14
LASSO$_a$	.2085	5
EBVS	.2078	1
EBVS$_i$	.2121	5

Presented in Table 3 are the 11 SNPs identified to have non-zero regression coefficients by adaptive lasso, EBVS, and EBVS$_i$. The only SNP, 102773, which was identified by EBVS, was identified by all other methods. While adaptive lasso and EBVS$_i$ each identified five SNPs with non-zero regression coefficients, there are only three commonly identified SNPs, i.e., 053887, 102773, and 065143. Both SNP 133907 identified by EBVS$_i$ and SNP 079089 identified by EBVS reside on chromosome 17, and are neighboring to each other with $16k$ bases in between. Instead 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 [2] to control FDR at 0.1, none of these methods reported any significant SNPs, though SNP

Table 3: Results of Analyzing Framingham Data.

		Chromosome-SNP						
		1-053887	4-510894	4-1361174	5-102773	8-065143	17-133907	17-079089
$\hat{\beta}$	LASSO	.0412	0	.0355	.0402	0	0	0
	LASSO_a	.1521	0	.0434	.1539	-.0200	0	.0167
	EBVS	0	0	0	.3778	0	0	0
	EBVS_i	.2417	-.0542	0	.3047	-.0857	.1093	0
p-value	LASSO	.2694	.1998	1	.6050	1	1	1
	LASSO_a	.2060	.0490	1	.0003	1	1	1
	EBVS	.3138	.1998	1	.0187	1	1	1
	EBVS_i	.0837	.1998	1	.0034	1	1	1
ζ	EBVS	.1277	.0133	.0347	.9976	.0981	.0869	.0966
	EBVS_i	.7609	.5275	.3269	.9718	.7464	.8450	.0009

102773 by adaptive lasso has the p-value as small as 0.0003. Instead, when controlling $\widehat{FDR}(\kappa) \leq 0.1$ for both EBVS and EBVS_i, EBVS identified only SNP 102773, and EBVS_i identified both SNP 102773 and 133907, with $\kappa = 0.8$. Note that SNP 133907 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

Here an empirical Bayes variable selection (EBVS) is proposed to extend empirical Bayes thresholding [19] 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 implement it by iteratively minimizing the conditional loss function (11). Without consideration of parallel computation, we can cycle through each coordinate of the parameters to minimize this loss function, which results in the algorithm described in (14). 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 [12], Daubechies et al. [8], Wu and Lange [34], and Breheny and Huang [5]. However, a direct application of the

coordinate descent algorithm to minimize the Bayes risk, or equivalently the conditional expectation (10), is challenged with the same difficulties as in directly minimizing the Bayes risk. However, an ICM/M algorithm can be easily implemented.

Without *a priori* information other than that regression coefficients are sparse, many lasso-type methods have been proposed with some tuning parameters. It is challenging 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 [26], the limited data usually contain large number of SNPs which differ in a tiny pool of individuals. It is almost infeasible to take a cross-validation method as the tiny pool of unique individuals for a rare variant is more likely to be included in the same fold. Instead, our proposed empirical Bayes variable selection obtains data-driven hyperparameters via conditional modes of the ICM/M algorithm, which takes full advantage of each precious 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 [33], and Meinshausen et al. [24] 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 lowers the power of variable selection and p-value calculation, especially for high-dimensional data of small sample sizes. Again, it is almost infeasible to apply such a splitting method to genome-wide association studies with rare variants.

As shown in Section 4, an Ising model as (23) 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) at one time, and therefore, many classical approximation methods for low-dimensional

issues may be used to simplify the implementation. On the other hand, the Ising prior (23) 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.

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 $\hat{\beta}_1^{(k+1)}$, $\cdots$, $\hat{\beta}_p^{(k+1)}$, $\hat{\sigma}^{(k)}$, and $\hat{\omega}^{(k)}$, 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 computed 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,

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

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

The conditional median $\hat{\beta}_j^{(k+1)}$ can be computed 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

[1] Maria M. Barbieri and James O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32:870–897, 2004.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

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

[4] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B*, 48:259–302, 1986.

[5] Patrick Breheny and Jian Huang. Cooridinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5:232–253, 2011.

[6] Bradley P. Carlin and Siddhartha Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B*, 57:473–484, 1995.

[7] L. Adrienne Cupples, Heather T. Arruda, Emelia J. Benjamin, and *et al.* The framingham heart study 100k snp genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics*, 8(Suppl1):S1, 2007.

[8] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[9] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[10] Sandrine Dudoit, Juliet P. Shaffer, , and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

[11] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[12] Wenjiang J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[13] Edward I. George and Robert E. McCulloch. Variable selection via gibbs sampling. *Journal of American Statistical Association*, 85:398–409, 1993.

[14] Charles J. Geyer. Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

[15] Charles J. Geyer and Elizabeth A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series B*, 54: 657–699, 1992.

[16] Xavier Guyon and Hans R. Kunsch. Asymptotic comparison of estimators in the ising model. In Piero Barone, Arnoldo Frigessi, and Mauro Piccioni, editors, *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, pages 177–198. Springer New York, 1992.

[17] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33:730–773, 2005.

[18] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of Landon Series A*, 196:453–461, 1946.

[19] Iain M. Johnstone and Bernard W. Silverman. Needles and straw in haystacks: empirical bayes estimates of possibly sparse sequence. *The Annals of Statistics*, 32:1594–1649, 2004.

[20] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The Annals of Applied Statistics*, 4:1498–1516, 2010.

[21] Fan Li and Nancy R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with application in genomics. *Journal of the American Statistical Association*, 105:1202–1214, 2010.

[22] Gang Liang and Bin Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE Transactions on Signal Processing*, 51:2043–2053, 2003.

[23] Shigeru Mase. Marked gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Mathematische Nachrichten*, 209:151–169, 2000.

[24] Nicolai Meinshausen, Lukas Meier, and Peter Buehlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681, 2009.

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

[26] Tal Nawy. Rare variants and the power of association. *Nature Methods*, 9:324, 2012.

[27] Michael A. Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176, 2004.

[28] Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65:117–149, 1943.

[29] Wei Pan, Benhuai Xie, and Xiaotong Shen. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66:474–484, 2010.

[30] Francesco C. Stingo, Yian A. Chen, Mahlet G. Tadesse, and Marina Vannucci. Incorporating biological information into linear models: a bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5:1978–2002, 2011.

[31] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B*, 58:267–288, 1996.

[32] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

[33] Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *Annals of Statistics*, 37:2178–2201, 2009.

[34] Tong T. Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.

[35] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society Series B*, 68:49–67, 2006.

[36] Xiang Zhou and Scott C. Schmidler. Bayesian parameter estimation in ising and potts models: a comparative study with applications to protein modeling. Technical report, Duke University, 2009.

[37] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.