Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1372–1382. doi:10.1111/biom.12882.

Scalable Bayesian Variable Selection for Structured High-dimensional Data

Changgee Chang^{1,*}, Suprateek Kundu^{2,**}, and Qi Long^{1,***}

¹Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, U.S.A.

²Department of Biostatistics, Emory University, Atlanta, GA, U.S.A.

Summary:

Variable selection for structured covariates lying on an underlying known graph is a problem motivated by practical applications, and has been a topic of increasing interest. However, most of the existing methods may not be scalable to high dimensional settings involving tens of thousands of variables lying on known pathways such as the case in genomics studies. We propose an adaptive Bayesian shrinkage approach which incorporates prior network information by smoothing the shrinkage parameters for connected variables in the graph, so that the corresponding coefficients have a similar degree of shrinkage. We fit our model via a computationally efficient expectation maximization algorithm which is scalable to high dimensional settings ($p \sim 100,000$). Theoretical properties for fixed as well as increasing dimensions are established, even when the number of variables increases faster than the sample size. We demonstrate the advantages of our approach in terms of variable selection, prediction, and computational scalability via a simulation study, and apply the method to a cancer genomics study.

Keywords

adaptive Bayesian shrinkage; EM algorithm; oracle property; selection consistency; structured high-dimensional variable selection

1. Introduction

With the advent of modern technology such as microarray analysis and next generation sequencing in genomics, recent studies rely on increasingly large amounts of data containing more than tens of thousands of variables. For example, in genomics studies, it is common to collect gene expressions from $p \approx 20,000$ genes, which is often considerably larger than the number of subjects in these studies, resulting in a classical small n large p problem. In

*changgee@pennmedicine.upenn.edu

**suprateek.kundu@emory.edu

***qlong@pennmedicine.upenn.edu

7. Supplementary Materials

Web Appendix A referenced in Section 2.1, Web Appendix B referenced in Sections 2.1 and 2.2, Web Appendix C referenced in Section 3, and Web Appendix D referenced in Section 4 are available with this paper at the Biometrics website on Wiley Online Library. The R package and the codes that reproduce the analysis results are available online.

addition, it is well-known that genes lie on a graph of pathways where nodes represent genes and edges represent functional interactions between genes and gene products. Currently, there exist several biological databases which store gene network information from previous studies (Kanehisa and Goto, 2000; Stingo et al., 2011), and these databases are constantly updated and augmented with newly emerging knowledge.

The classical variable selection approaches such as Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), spike and slab methods (Mitchell and Beauchamp, 1988; George and McCulloch, 1993), and many of their derivatives have been successful, but they are limited in that they do not exploit the association structure between variables. In fact, when genes are known to lie on an underlying graph, it is a common practice to treat connected covariates as a group, especially in the light of increasing evidence that incorporating such prior information can produce biologically meaningful outcomes, and lead to improvements in prediction and variable selection in analysis of high dimensional data (Li and Li, 2008; Pan et al., 2010; Stingo et al., 2011; Stingo and Vannucci, 2011).

Existing approaches incorporating graph structured covariates focus on network based penalties in the frequentist paradigm, and network informed priors in the Bayesian paradigm. Some of the frequentist approaches encourage connected covariates to have similar effect sizes. For example, Li and Li (2008) and Pan et al. (2010) proposed network-based penalties, which induced sparsity of estimated effects while encouraging similar magnitude of effects for connected variables. In the Bayesian framework, Li and Zhang (2010), Stingo and Vannucci (2011), and Stingo et al. (2011), used the spike and slab priors in combination of the Markov random field (MRF) prior to incorporate graph information. Recently, Rockova and George (2014) proposed an expectation maximization (EM) algorithm for a Bayesian model, which was implemented via a variational approximation.

Unfortunately, these existing methodologies, while promising, are beset with one or more limitations. The Bayesian approaches involving the MRF priors are implemented using Markov Chain Monte Carlo (MCMC) and hence are not scalable up to tens of thousands of variables, such as in our cancer genomics application. Moreover, the MRF prior is subject to the phase transition problem, which results in adjacent variables often being either highly correlated or hardly correlated. On the other hand, the network based penalty approaches (Li and Li, 2008; Pan et al., 2010) may be too restrictive in smoothing over the magnitudes of effect sizes for connected variables. Instead, we propose and demonstrate that it is more flexible to smooth over the shrinkage parameters for the regression coefficients of connected variables which enforces adaptive shrinkage, but does not place any restriction on the effect sizes of connected variables. In addition, the motivation of our work is on developing a scalable structured variable selection approach which has desirable theoretical and numerical properties for ultra high dimensional settings where the dimension increases much faster than the sample size.

We propose a Bayesian shrinkage approach and an associated EM algorithm for variable selection with structured covariates, which assigns independent Laplace priors on the regression coefficients, while incorporating the underlying graph information via the hyperprior imposed on the shrinkage parameters of the Laplace distributions. Specifically,

the shrinkage parameters are assigned a log-normal prior with its inverse covariance matrix having a graph Laplacian structure (Chung, 1997), which specifies zero or positive partial correlations for pairs of disconnected and connected variables, respectively. The graph Laplacian structure smooths the shrinkage parameters of connected variables, while allowing the shrinkage parameters of disconnected variables to be conditionally independent, and also guarantees the existence of a *maximum-a-posteriori* (MAP) estimator as elaborated in the sequel. The proposed approach enables adaptive smoothing of the shrinkage parameters corresponding to adjacent variables in the underlying graph which encourages pairs of connected important(unimportant) variables to be jointly selected(excluded) in the model without requiring the coefficients to be numerically similar. It also allows the shrinkage parameters for connected pairs of important and unimportant variables to be weakly conditionally correlated, thereby enabling the model to weight these connected variables differently.

Although the proposed model can be implemented using MCMC, it is not scalable to high dimensional settings of our interest. Moreover, MCMC samples cannot take exact zeroes under a Laplace prior, which could be useful for model selection. We bypass this limitation via an EM algorithm to obtain the MAP estimate, where we treat the inverse covariance matrix in the log-normal hyperprior on the shrinkage parameters as missing variables, and implicitly marginalize them over. We employ recent computational developments such as the dynamic weighted Lasso (DWL) (Chang and Tsay, 2010) to speed up each EM iteration, which leads to scalability to ultra-high dimensional settings.

We note that the idea of smoothing over shrinkage parameters based on prior pathway knowledge has been previously proposed by Rockova and Lesaffre (2014) who incorporated pathway membership information via Gamma priors on the shrinkage parameters in contrast to the proposed approach which models these parameters based on more refined edge level information via log-normal priors. Hence unlike the proposed approach, the Rockova and Lesaffre (2014) method is not able to distinguish between prior network knowledge having the same membership information but different edge structures, which may constrain its performance. Moreover, there are implicit differences in the implementation of the EM algorithm between the two approaches. By carefully designing an algorithm which treats the covariance structure of the shrinkage parameters as missing variables, we are able to scale up to very high dimensions which may not be practical under Rockova and Lesaffre (2014), according to the discussions by the authors. Finally, the proposed approach possesses consistency and oracle properties for more flexible settings involving dimensions growing much faster than the sample size.

We also investigate the connections between our proposed model and the adaptive Lasso (Zou, 2006) and show that our EM estimator can be expressed as an adaptive Lasso type estimator. However, it is noted that the classical adaptive Lasso simply computes the shrinkage parameters via a data-driven approach without taking into account the correlation structure between covariates. In contrast, our approach adds an additional level of hierarchy to the generic adaptive lasso approach, by specifying network informed priors on the coefficient specific shrinkage parameters, so that the shrinkage parameters are estimated adaptively with respect to the effect sizes and prior graph knowledge. Another appealing

feature is that the proposed estimator retains the oracle property demonstrated by the adaptive Lasso type of estimators (Fan and Li, 2001; Zou, 2006; Huang et al., 2008a,b; Armagan et al., 2013), even when the graph information is mis-specified. Our simulation study shows that the proposed estimator outperforms Lasso based and spike and slab type approaches when the graph information is correctly specified or partially mis-specified. Moreover, a variant of the proposed method which does not require prior knowledge provides more accurate answers compared to traditional penalized approaches assuming independence between covariates.

The rest of this article is organized as follows. We present the proposed methodology and the EM algorithm in Section 2, the theoretical results in Section 3, and the simulation results comparing our approach with several competitors in Section 4. We apply our method to a cancer genomics study in Section 5.

2. Methodology

2.1 Model Specification

Let $\mathbf{0}_m$ and $\mathbf{1}_m$ denote the length- m vectors with 0 entries and 1 entries, respectively, and I_m the $m \times m$ identity matrix. The subscript m may be omitted in the absence of ambiguity. For any length- m vector $\mathbf{v}$, we define $e^{\mathbf{v}} = (e^{v_1}, \dots, e^{v_m})'$, $\log \mathbf{v} = (\log v_1, \dots, \log v_m)'$, $|\mathbf{v}| = (|v_1|, \dots, |v_m|)'$, and $D_{\mathbf{v}} = \text{diag}(\mathbf{v})$.

Suppose we have a random sample of n observations $\{y_i, \mathbf{x}_i; i=1, \dots, n\}$ where y_i is the outcome variable and $\mathbf{x}_i$ is a vector of p predictors. Let $\mathcal{G} = \langle V, E \rangle$ denote the known underlying graph for the p predictors, where $V = \{1, \dots, p\}$ is the set of nodes and $E \subset \{(j, k) : 1 \leq j < k \leq p\}$ is the set of undirected edges. Let G be the $p \times p$ adjacency matrix in which the (j, k) -th element $G_{jk} = 1$ if there is an edge between predictors j and k , and $G_{jk} = 0$ otherwise.

Consider the linear model

$$\mathbf{y} = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad (1)$$

where $\mathbf{y} = (y_1, \dots, y_n)'$, $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, $\beta = (\beta_1, \dots, \beta_p)'$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$, and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. We assign the following priors to β and σ^2

$$\beta_j \sim \mathcal{L}(\lambda_j / \sigma), \sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma), \quad j = 1, \dots, p, \quad (2)$$

where λ_j is the shrinkage parameter for β_j , $\mathcal{L}(\cdot)$, and $\mathcal{IG}(\cdot)$ denote the Laplace and inverse gamma distributions, respectively. The prior in (2) looks similar to Bayesian Lasso (Park and Casella, 2008), but we treat the shrinkage parameters λ_j as random variables encoding the graph knowledge $\mathcal{G}$ under an informative network based prior. In particular, we specify

$$\boldsymbol{\alpha} = (\log(\lambda_1), \dots, \log(\lambda_p))' \sim \mathcal{N}(\boldsymbol{\mu}, \nu \boldsymbol{\Omega}^{-1}), \quad (3)$$

where $\Omega(j, k) = \Omega(k, j) = -\omega_{jk} = -\omega_{kj}$ for $1 \leq j < k \leq p$, and $\Omega(j, j) = 1 + \sum_{k: j < k \leq p} \omega_{jk}$ for $1 \leq j < k \leq p$, and we assign the following prior to $\boldsymbol{\omega} = \{\omega_{jk} : j < k\}$

$$\pi(\boldsymbol{\omega}) \propto \left| \boldsymbol{\Omega} \right|^{-1/2} \prod_{G_{jk}=1} \omega_{jk}^{a_w-1} \exp(-b_w \omega_{jk}) 1(\omega_{jk} > 0) \prod_{G_{jk}=0} \delta_0(\omega_{jk}), \quad (4)$$

where δ_0 is the Dirac delta function concentrated at 0 and $1(\cdot)$ is the indicator function.

The diagonally dominant structure of $\boldsymbol{\Omega}$ specified above ensures positive definiteness. Note that the constant 1 in the diagonal entries of $\boldsymbol{\Omega}$ prevents the identifiability problem between ν and other elements in $\boldsymbol{\Omega}$, and causes no additional restriction in the model. According to (4), the partial correlations among the log-shrinkage parameters corresponding to connected variables are only positive, and it is zero otherwise. This feature enables us to implement network based shrinkage of regression coefficients by (a) smoothing the shrinkage parameters λ_j and λ_k via positive partial correlations if predictors j and k are connected, which encourages a similar degree of shrinkage for the corresponding coefficients; and (b) allowing λ_j and λ_k to be conditionally independent for unconnected variables, so that the corresponding regression coefficients are not constrained to rely on each other. Proposition 1 shows that the prior in (4) is proper. The proof is provided in Web Appendix A.

PROPOSITION 1: The prior $\pi(\boldsymbol{\omega})$ of $\boldsymbol{\omega}$ in (4) is proper.

We note that Liu et al. (2014) used a similar graph Laplacian matrix to model the inverse covariance matrix of the regression coefficients; however their method did not rely on prior graph knowledge, unlike the proposed approach which assigns zero off-diagonal elements in the graph Laplacian structure for the inverse covariance matrix of the log-shrinkage parameters. Moreover, their method resulted in a OSCAR type shrinkage (Bondell and Reich, 2008), whereas, the proposed approach adaptively smooths the regression coefficients for connected variables and results in non-convex shrinkage for $\nu > 0$, as elaborated in the sequel.

Our model formulation has several appealing features. First, a higher positive partial correlation between two connected predictors results in the smoothing of the shrinkage parameters which leads to an increased probability of including or excluding both predictors simultaneously under an EM algorithm. This feature is helpful when both variables are important or unimportant but may have different effect sizes, which is usually the case in practical applications. Second, in the scenario where one of the paired predictors is important but the other one is not, the method can learn from the data and impose a weak partial correlation, thereby enabling the corresponding shrinkage parameters to act in a largely uncorrelated manner (refer Figure 3 and Section 4.2 for numerical evidence). Third,

the selection of unconnected variables is guided by partially uncorrelated shrinkage parameters, and the corresponding effect sizes are estimated independently of each other.

The operating characteristics of the model are controlled by tuning parameters μ and ν in (3). The mean vector μ determines the locations of α and controls the overall sparsity of the model, with higher values translating to a sparser model. A default choice would be $\mu_j = \mu$ for some $\mu \in \mathbb{R}$ unless prior evidence suggests otherwise. Figure 1(a) plots the marginal density of the regression coefficients for different values of μ with λ marginalized out (via Monte Carlo averaging), while ν and σ are kept fixed. It is clear that larger μ values lead to sharper peaks at zero with lighter tails, thus encouraging greater shrinkage. Moreover, ν specifies the prior confidence on the choice of μ , and also controls the type of shrinkage (convex or concave). In particular, small values of ν lead to $\alpha \approx \mu$, and in the extreme case when $\nu = 0$, we have $\alpha = \mu$, resulting in a Lasso type shrinkage for fixed μ , or an adaptive lasso shrinkage when μ is chosen in a data adaptive manner. The role of ν in regulating shrinkage is evident from Figures 1(b)/1(d), which plots the density/negative logarithm of the density for the marginal regression coefficients for different values of ν , while μ and σ are fixed. Figures 1(b) and 1(d) also show that larger values of ν result in higher-peaked and heavier-tailed densities than Lasso and the corresponding penalty becomes similar to non-convex penalties in the frequentist literature, e.g. SCAD in Fan and Li (2001), while smaller values lead to Lasso type shrinkage. For our applications, we propose a strictly positive value of ν which works well in a variety of scenarios, and treat μ as a fixed tuning parameter.

The magnitudes and variability of the partial correlations are influenced by the shape parameter a_ω and the rate parameter b_ω in (4), which serve the similar roles as those of the gamma distribution. To see how they affect the correlations, consider $p = 2$ and $G_{12} = 1$. It follows that the joint prior density of α_1 and α_2 after marginalizing out ω_{12} is given (up to a

$$\text{constant) by } \pi(\alpha_1, \alpha_2) \propto f(\alpha_1, \alpha_2) = \exp\left(-\frac{(\alpha_1 - \mu_1)^2 + (\alpha_2 - \mu_2)^2}{2\nu}\right) \left(b_\omega + \frac{(\alpha_1 - \alpha_2)^2}{2\nu}\right)^{-a_\omega} \quad \text{Figure 2}$$

draws the contour plots of $f(\alpha_1, \alpha_2)$ for 4 different combination of a_ω and b_ω ; $(a_\omega, b_\omega) = (1, 1), (1, 4), (4, 1), (4, 4)$ with $\mu_1 = \mu_2 = 1$ and $\nu = 1$. As a_ω increases and/or b_ω decreases, α_1 and α_2 tend to have a stronger correlation, translating to a higher probability of having similar values. This is also evident in the E-step in the EM algorithm (see equation (2) in Web Appendix B), where large values of a_ω/b_ω tend to result in large expected values of ω_{jk} , which in turn tends to result in similar values for $\alpha_j - \mu_j$ and $\alpha_k - \mu_k$.

When a small proportion of important and unimportant variables are connected, then a stronger smoothing can enhance the performance under our approach, which can be achieved via a large prior mean and precision (large a_ω , small b_ω). On the other hand, if one prefers that the data inform the degree of smoothness, then a large variance on ω is preferable. This non-informative choice enables the algorithm to adaptively assign small values for the edges between important and unimportant variables, and to still encourage smoothness on the other edges. Our experience suggests that $2 \leq a_\omega \leq 4$ and $b_\omega = 1$ should work for a broad range of scenarios, although more general choices are also possible.

2.2 EM Algorithm

The MAP estimator for the proposed model is obtained by maximizing the posterior density with ω marginalized out, i.e. $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2, \hat{\alpha}) = \operatorname{argmax}_{\theta} \int \pi(\theta, \omega | y, X) d\omega$. It is possible to demonstrate that the posterior density for θ is given by

$$\pi(\theta | y, X) \propto \pi(y | \beta, \sigma^2, X) \pi(\beta | \sigma^2, \alpha) \pi(\sigma^2) \times \exp\left(-\frac{(\alpha - \mu)'(\alpha - \mu)}{2v}\right) \quad (5)$$

$$\times \prod_{j < k, G_{jk} = 1} \left(b_{\omega} + \frac{1}{2v}(\alpha_j - \alpha_k)^2\right)^{-a_{\omega}}.$$

We note that since the logarithm of marginal posterior density may not be convex, the MAP estimator may have multiple (local) solutions. However, extensive numerical studies illustrate that the performance of our EM estimator is robust to this problem, and the algorithm overcomes the local solution issue for large sample sizes by admitting a unique solution asymptotically; see Section 3.

Although one can directly optimize the objective function to compute $\hat{\theta}$, we choose to use the EM algorithm to obtain the MAP estimate by treating ω as the missing variable. This is done to exploit a computational advantage of the fact that the Hessian matrix with respect to α is guaranteed to be positive definite under the EM algorithm, which is not necessarily the case if one directly optimizes the marginal posterior density $\pi(\theta | y, X)$ with respect to θ . Since the EM algorithm exploits part of the curvature information in optimizing with respect to α at nearly no extra computational cost, it is expected to lead to a reduced number of iterations and savings in computation time. In the absence of graph information ($\Omega = I_p$), we call the resulting estimator *EM* estimator for Bayesian *SH*rinkage approach, or EMSH in short. In the presence of graph information, we call it EMSH with the *Structural* information incorporated, or EMSHS in short.

2.3 Role of Shrinkage Parameters

It is straightforward to show that EMSHS estimator satisfies

$$\hat{\beta} = \operatorname{argmin}_{\beta} \frac{1}{2}(y - X\beta)'(y - X\beta) + \sum_{j=1}^p \hat{\sigma} e^{\hat{\alpha}_j} |\beta_j|, \quad (6)$$

$$e^{\hat{\alpha}_j} |\hat{\beta}_j| = \frac{\hat{\sigma}}{v} (\mu + v - \hat{\alpha}_j + \sum \omega_{jk}^{(\infty)} (\hat{\alpha}_k - \hat{\alpha}_j)), \quad (7)$$

where $\Omega^{(\infty)}$ is the final value of Ω from the EM algorithm. When $\hat{\sigma}$ and $\hat{\alpha}$ are fixed, the solution $\hat{\beta}$ in (6) resembles an adaptive lasso solution with the regularization parameter

$\hat{\xi} = \hat{\sigma}e^{\hat{\alpha}} \cdot \omega_{jk}^{(\infty)}$ suggests that larger values of $\hat{\alpha}_j$ translate to smaller values for $|\hat{\beta}_j|$ and vice-versa, clearly demonstrating the role of the shrinkage parameters α .

Unlike the original adaptive lasso that uses the fixed weights, EMSHS iteratively adjusts the weights based on the current coefficients and the underlying graph knowledge. By estimating the weights in such an adaptive manner, EMSHS can avoid having to specify an initial consistent estimator for the weights as in the original adaptive lasso. This advantage is essentially similar to that of the iterated lasso (Huang et al., 2008b). However, unlike the iterated lasso approach, the weights of EMSHS also reflects the structural information. Equation (7) shows that, in the presence of the graphical information, the shrinkage parameter for β_j also depends on the other shrinkage parameters for the variables adjacent to x_j and the corresponding partial correlations. The adaptive learning of the smoothness between shrinkage parameters based on prior network knowledge is the key distinction of our proposed Bayesian method compared to the iterated lasso approach which yields superior numerical performance. Please refer to Figure 3 and Section 4.2 for numerical evidence and further discussions.

3. Theoretical Properties

Here, we establish that the proposed EMSHS estimator has desirable theoretical properties. To fix ideas, let p_n denote the number of candidate predictors, and q_n denote the true model size important, both of which vary with the sample size n . Model (1) is rewritten as

$$y_n = X_n \beta_0 + \epsilon_n,$$

where y_n is the $n \times 1$ response vector, X_n is the $n \times p_n$ design matrix, β_0 is the $p_n \times 1$ true coefficient vector, and ϵ_n is the $n \times 1$ error vector. The errors are independent Gaussian with mean 0 and variance σ_0^2 and independent of the covariates. X_n is standardized such that $\mathbf{1}' \mathbf{x}_{nj} = 0$ and $\mathbf{x}_{nj}' \mathbf{x}_{nj} = n$ for $j = 1, \dots, p_n$, where $\mathbf{x}_{nj}$ is the j -th column (variable) of X_n , and let $\sum_n = \frac{1}{n} X_n' X_n$ be the sample covariance matrix.

Let $\hat{\theta}_n = (\hat{\beta}_n', \hat{\sigma}_n^2, \hat{\alpha}_n')'$ be the EMSHS solution. Let $\mathcal{A}_n = \{j: \hat{\beta}_{nj} \neq 0\}$ be the index set of the selected variables in $\hat{\beta}_n$, and $\mathcal{A}_n = \{j: \hat{\beta}_{0j} \neq 0\}$ be the index set of the true important variables. For any index set $\mathcal{A}$, $\mathbf{v}_{\mathcal{A}}$ represents the subvector of a vector $\mathbf{v}$ with entries corresponding to $\mathcal{A}$. $E_{\mathcal{A}\mathcal{B}}$ is the submatrix of a matrix E with rows and columns corresponding to $\mathcal{A}$ and $\mathcal{B}$, respectively. When a sequential index set $\mathcal{A}_n$ is used for a sequence of vectors or matrices indexed by n , the subscript n may be omitted for conciseness if it does not cause a confusion. For example, $\mathbf{v}_{n\mathcal{A}_n}$ can be written as $\mathbf{v}_{\mathcal{A}_n}$ or $\mathbf{v}_{n\mathcal{A}}$, and $E_{n\mathcal{A}_n\mathcal{B}_n}$ as $E_{\mathcal{A}_n\mathcal{B}_n}$ or $E_{n\mathcal{A}\mathcal{B}}$. Finally, $\rightarrow_p$ and $\rightarrow_d$ denote convergence in probability and in distribution, respectively. The following result describes the asymptotic properties of the proposed approach under certain standard regularity conditions.

THEOREM 1: Under certain regularity conditions listed in Web Appendix C, the following statements hold for the EMSHS estimator $\hat{\theta}_n$ as $n \rightarrow \infty$ when $p_n = O(\exp(n^U))$, $q_n = O(n^u)$, $0 \leq U < 1$, $0 \leq u < (1 - U)/2$, and $q_n \leq p_n$.

$$(1) \quad P(\mathcal{A}_n = \mathcal{A}_0) \rightarrow 1.$$

$$(2) \quad \text{Letting } s_n^2 = \gamma'_n \sum_{n \in \mathcal{A}_0}^{-1} \gamma_n \text{ for any sequence of } q_n \times 1 \text{ nonzero vectors } \gamma_n, \text{ we have}$$

$$n^{1/2} s_n^{-1} \gamma'_n (\hat{\beta}_{n \mathcal{A}_0} - \hat{\beta}_{0 \mathcal{A}_0}) \rightarrow_d \mathcal{N}(0, \sigma_0^2).$$

$$(3) \quad \text{The solution is unique with probability tending to 1 as } n \rightarrow \infty.$$

The conditions and the proofs are provided in Web Appendix C, which also contains the details for the oracle property under fixed p scenarios. We note that the above result holds even when the prior graph structure is mis-specified, as elaborated in Web Appendix C.

4. Simulation Study

We conduct simulations to evaluate the performance of the proposed approach in comparison with several existing methods. The competing methods include the lasso (Lasso), the adaptive Lasso (ALasso), the group Lasso (GLasso) (Yuan and Lin, 2006; Zeng and Breheny, 2016), the network-constrained regularization approach (Net) (Li and Li, 2008), the Bayesian variable selection approach using spike and slab priors and MRF priors by Stingo et al. (2011) which we denote as BVS-MRF, and finally the EM approach for Bayesian variable selection (denoted as EMVS) proposed by Rockova and George (2014) and its extension to incorporate structural information (denoted as EMVSS). Among these approaches, GLasso treats pathways as overlapping groups, while Net, EMSHS, and EMVSS incorporate the graph information, and BVS-MRF exploits both. For Lasso and ALasso, we use the glmnet R package where the initial consistent estimator for ALasso is given by the ridge regression. We used the R packages grpregOverlap (Zeng and Breheny, 2016) and glmgraph (Chen et al., 2015) for GLasso and Net, respectively. The Matlab code for BVS-MRF is provided with the original article by Stingo et al. (2011). The unpublished R codes for EMVS and EMVSS were provided to us by the authors.

4.1 Simulation Set-up

The simulated data are generated from the following model

$$y_i = x_i' \beta + \epsilon_i, \quad 1 \leq i \leq n,$$

where $x_i \sim \mathcal{N}\left(0, \sum_X\right)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The number q of nonzero coefficients in β is randomly chosen between 4 and 8, and the true non-zero effect sizes were randomly

generated from a $\mathcal{U}(0.5, 2)$ distribution. The sample size is fixed at $n = 100$, the residual variance is fixed at $\sigma_e^2 = 2$, and we consider $p = 1,000, 10,000$, and $100,000$.

Let G_0 be the adjacency matrix for the true graph that governs Σ_X , where $G_{0,jk} = 1$ if there is an edge between predictors j and k , and $G_{0,jk} = 0$ otherwise. The mechanism to generate G_0 and Σ_X is described in Web Appendix D, which resembles practical gene networks encountered in applications, and where a subset of important and unimportant genes lie in the same pathway and are connected. We use a working graph G given to fit the models, where the working graph may be specified correctly or mis-specified, as follows.

1. $G = G_0$.
2. $G = G_0$, where G_0 does not allow edges between important and unimportant variables.
3. G_0 is as in scenario 1, but G is randomly generated with the same number of edges as G_0 .
4. G_0 is as in scenario 2, but G is randomly generated with the same number of edges as G_0 .
5. G_0 is as in scenario 1, but G includes only a subset of the edges in G_0 for which the corresponding partial correlations are greater than 0.12.

Scenarios 1 and 2 are the cases where the true graph is completely known; scenario 2 allows no correlation between important and unimportant variables and hence is an ideal setting for approaches which encourage connected variables to be jointly included or excluded in the model. Scenarios 3 and 4 considered as the worst case scenarios, since G is completely mis-specified. Scenario 5 mimics the situation where only strong signals from G_0 are known to data analysts, which is intermediate between the ideal and the worst case scenarios.

Although each of the five scenarios consist of a distinct edge structure for G , scenarios 1, 3, and 5 have the same pathway membership information, and similarly for scenarios 2 and 4.

We generate 500 data sets, each of which contains the training set, the validation set, and the test set of size $n = 100$ each. Variable selection performance is assessed in terms of the rates of false positives (FP) and false negatives (FN), and the prediction performance is assessed in terms of mean squared prediction error (MSPE) for the test data. Note that it is a paired design, which means all the methods were trained, validated, and tested with the same data, and thus we report the standard error of the differences in MSPE. We also report the average computation time per tuning parameter value in seconds. For the proposed approach, we choose priors $\sigma^2 \sim \mathcal{IG}(1, 1)$ and $(a_\omega, b_\omega) = (2, 1)$ for ω_{jk} , both of which are fairly uninformative. The remaining parameters μ and ν are chosen by validation method.

4.2 Results

The simulation results are summarized in Table 1. Note that Net, EMVS, and EMVSS were not feasible for $p = 100,000$ and BVS-MRF was not feasible for $p = 10,000$, due to either exceedingly large running time or lack of memory.

When true graph knowledge is available, the structured variable selection methods EMVSS and EMSHS have superior prediction performance compared to their counterparts that do not use graph information (i.e. EMVS and EMSH). Moreover, the performance of each structured variable selection method (namely, EMSHS, EMVSS and BVS-MRF) improves from scenario 3 to scenario 1, as well as from scenario 4 to scenario 2, further demonstrating the benefits of the correctly specified graph information. Similarly, the partially correctly specified graph information in Scenario 5 also improves the performance of these methods. The prediction error under Li and Li (2008), although higher than the proposed approach, improves from scenario 4 to scenario 2, but not from scenario 3 to scenarios 1 and 5. In contrast, we note that the GLasso reports identical prediction performance under Scenarios 1,3,5, as well as Scenarios 2 and 4, and hence is unable to distinguish between these cases which have the similar pathway membership information but different edge structures. We also note that the improvement in the prediction under EMSHS compared to EMVSS and NET increases with the dimension of the feature space, which suggests a distinct advantage in high dimensions.

When the working graph G is partially or fully correctly specified (scenarios 1, 2, and 5), EMSHS significantly outperforms all the other methods in terms of prediction, with the relevance of significance being captured via the difference in MSE reported in parenthesis in Table 1. It is remarkable that while EMSH has a superior prediction performance when the graph information is completely mis-specified, EMSHS still yields close to the best prediction performance, demonstrating its robustness to mis-specified graph information. The robustness comes from the ability to adaptively learn the correlation between shrinkage parameters. In particular, the proposed method learns small values of the partial correlations between pairs of connected important and unimportant variables resulting in weak smoothing, and imposes stronger partial correlations for other sets of connected variables, which enable accurate variable selection and prediction (see Figure 3).

We also see that Lasso, ALasso, GLasso, Net, and BVS-MRF suffer significantly high FP rates in all scenarios, suggesting an inflated estimated model. On the other hand, BVS-MRF has significantly high FN rates and EMVS has high FN rates in Scenario 1. In contrast, EMSHS typically has the lowest FP rate among all approaches, and exhibits a FN rate which is the lowest or second lowest among all approaches incorporating prior knowledge when it is correctly or partially correctly specified. Finally, although somewhat slower than the Lasso and adaptive Lasso, the proposed approach is still computationally efficient and is scalable up to $p = 100,000$ and higher dimensions. In contrast none of the competing approaches incorporating prior graph knowledge except the group lasso is scalable to such high dimensions.

5. Data Application

We applied the proposed method to analysis of a glioblastoma data set obtained from the Cancer Genome Atlas Network (Verhaak et al., 2010). The data set includes the survival times (T) and gene expression levels for $p = 12,999$ genes (X) for 303 glioblastoma patients, and also the clinical characteristics (Z) for the patients. As glioblastoma is known as one of the most aggressive cancers, only 12% of the samples were censored.

We fit an accelerated failure time model using the uncensored data weighted by the inverse probability of not being censored (Johnson et al., 2011) $\log T_i = \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i$, $i = 1, \dots, n$, where ϵ_i 's are independent Gaussian random variables and all covariates were standardized to have mean 0 and variance 1. The censoring probability was estimated using a proportional hazard Cox model using Z as covariates. The network information ($\mathcal{G}$) for X was retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database including a total of 332 KEGG pathways and 31,700 edges in these pathways.

In addition to EMSHS and ESMH, we included several competing methods that are computationally feasible, namely, Lasso, ALasso, GLasso, Net, EMVS, and EMVSS. The optimal tuning parameters μ and ν were chosen by minimizing the 5-fold cross-validated mean-squared prediction error. We used $a_\sigma = 1$ and $b_\sigma = 1$ for prior of σ^2 , which is uninformative. As shown in Table 2, ESMH and EMSHS achieve the best prediction performance and both are substantially less expensive than EMVS and EMVSS in terms of computation. Similar to our simulation results, ESMH again yields better prediction performance than lasso and adaptive lasso, demonstrating the advantage of using the data to learn adaptive shrinkage in ESMH.

To assess variable selection behavior of EMSHS and ESMH, we conducted a separate experiment where we randomly divided the entire sample into two subsets; the first subset with 178 subjects (70% of the whole sample) was used as the training data and the second subset with 30% of the subjects was used as the validation data to select optimal tuning parameter values. We repeated this procedure 100 times. Of the 100 random splits, 20 genes were selected at least 5 times by ESMH and 24 genes were selected at least 5 times by EMSHS. Further examination reveals that a set of 6 genes that have been selected at least once by EMSHS but never by ESMH had edges with variables that have been selected by ESMH. On the other hand, a set of 18 genes that have been selected at least once by ESMH but never by EMSHS had edges with variables that have never been selected by ESMH. This lends support to the notion that incorporating graph information may reduce false positives and false negatives, which is consistent with the findings in our simulation study. In addition, the number of genes selected by several existing methods such as Lasso (20), GLasso (19) and Net (14) are comparable to or less than the number of genes selected by ESMH and EMSHS. Fourteen of the 20 genes selected by ESMH were also chosen by at least one other existing method considered, and seventeen of the 24 genes selected by EMSHS were also chosen by at least one other existing method considered.

Using the set of 24 genes selected by EMSHS, a gene list enrichment analysis conducted via the ToppGene Suite (Chen et al., 2009) identified a number of enriched pathways including, for example, the vitamin D pathway, which has been suggested to have an etiologic role in tumor development (Holick et al., 2007). In addition, several of these 24 genes have been associated with brain tumor. For example, Tang et al. (2014) showed that BRD7 is highly expressed in gliomas; Lee et al. (2010) reported that RANBP17 can enhance transcript levels of endogenous p21^{Waf1/Cip1}, a well-validated E2A transactivation target gene; and Sirvent et al. (2012) found that TOM1L1 depletion has been shown to decrease tumor growth in

xenografted nude mice. The signs of the regression coefficients for TOM1L1, RANBP17 and BRD17 are consistent with the existing knowledge for the genes in promoting/suppressing the development of cancer. Our data analyses demonstrate that EMSHS yields biologically meaningful results.

6. Discussion

This article introduces a scalable Bayesian regularized regression approach and an associated EM algorithm which can incorporate the structural information between covariates in high dimensional settings. The approach relies on specifying informative priors on the log-shrinkage parameters of the Laplace priors on the regression coefficients, which results in adaptive regularization. The method does not rely on initial estimates for weights as in adaptive lasso approaches, which provides computational advantages in higher dimensions as demonstrated in simulations. Appealing theoretical properties for both fixed and diverging dimensions are established, under very general assumptions, even when the true graph is mis-specified. The method demonstrates encouraging numerical performance in terms of scalability, prediction and variable selection, with significant gains when the prior graph is correctly specified, and a robust performance under prior graph mis-specification.

Although the proposed EM algorithm is scalable to very high dimensions, it does not enable quantification of uncertainty, which is a limitation of our approach. However, we note that our model allows a straightforward implementation of an MCMC algorithm via a MH-within-Gibbs approach which should be feasible for moderate to large dimensions. For very high dimensions where the MCMC may not work well, one can compute bootstrap confidence intervals by repeatedly fitting the proposed EM algorithm to multiple bootstrap samples. As another alternative, one could potentially adapt the sandwich estimator proposed in Fan and Li (2001) to obtain standard errors for the non-zero regression coefficients.

Extending the current approach to more general types of outcomes such as binary or categorical should be possible (McCullagh and Nelder, 1989), although the complexity of the optimization problem may increase. These issues can potentially be addressed using a variety of recent advances in literature involving EM approaches via latent variables (Polson et al., 2013), coordinate descent method (Wu and Lange, 2008), and other optimization algorithms. Another potential avenue is to extend the approach to more general classes of priors on the shrinkage parameters, which will translate to more diverse penalties. We hope to tackle these issues as future research questions of interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is partly supported by NIH/NCI grants (R03CA173770, R03CA183006, P30CA016520, and R01DK108070). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank Prof. David Dunson for helpful discussions and comments that led to

the motivation of this work. The authors also appreciate the editor comments which lead to considerable improvements.

References

- Armagan A, Dunson DB, and Lee J (2013). Generalized Double Pareto Shrinkage. *Statistica Sinica* 23, 119–143. [PubMed: 24478567]
- Bondell HD and Reich BJ (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* 64, 115–123. [PubMed: 17608783]
- Chang C and Tsay RS (2010). Estimation of Covariance Matrix via the Sparse Cholesky Factor with Lasso. *Journal of Statistical Planning and Inference* 140, 3858–3873.
- Chen J, Bardes EE, Aronow BJ, and Jegga AG (2009). ToppGene Suite for Gene List Enrichment Analysis and Candidate Gene Prioritization. *Nucleic Acids Research* 37, W305–W311. [PubMed: 19465376]
- Chen L, Liu H, Kocher JPA, Li H, and Chen J (2015). Glmgraph: An R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* 31, 3991–3993. [PubMed: 26315909]
- Chung FR (1997). Spectral graph theory, volume 92 American Mathematical Soc.
- Fan J and Li R (2001). Variable Selection via Nonconcave Penalized. *Journal of the American Statistical Association* 96, 1348–1360.
- George EI and McCulloch RE (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* 88, 881–889.
- Holick CN, Stanford JL, Kwon EM, Ostrander EA, Nejentsev S, and Peters U (2007). Comprehensive association analysis of the vitamin D pathway genes, VDR, CYP27B1, and CYP24A1, in prostate cancer. *Cancer Epidemiology Biomarkers and Prevention* 16, 1990–1999.
- Huang J, Ma S, and Zhang C-H (2008a). Adaptive Lasso for Sparse High-Dimensional Regression Models. *Statistica Sinica* 18, 1603–1618.
- Huang J, Ma S, and Zhang C-H (2008b). The iterated lasso for high-dimensional logistic regression. Technical report No. 392, The University of Iowa.
- Johnson BA, Long Q, and Chung M (2011). On path restoration for censored outcomes. *Biometrics* 67, 1379–1388. [PubMed: 21457193]
- Kanehisa M and Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *ucleic Acids Research* 28, 27–30.
- Lee J-H, Zhou S, and Smas CM (2010). Identification of RANBP16 and RANBP17 as Novel Interaction Partners for the bHLH Transcription Factor E12. *Journal of Cellular Biochemistry* 111, 195–206. [PubMed: 20503194]
- Li C and Li H (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24, 1175–1182. [PubMed: 18310618]
- Li F and Zhang NR (2010). Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics. *Journal of the American Statistical Association* 105, 1202–1214.
- Liu F, Chakraborty S, Li F, Liu Y, and Lozano AC (2014). Bayesian Regularization via Graph Laplacian. *Bayesian Analysis* 9, 449–474.
- McCullagh P and Nelder JA (1989). Generalized linear models, volume 37 CRC press.
- Mitchell TJ and Beauchamp JJ (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* 83, 1023–1032.
- Pan W, Xie B, and Shen X (2010). Incorporating Predictor Network in Penalized Regression with Application to Microarray Data. *Biometrics* 66, 474–484. [PubMed: 19645699]
- Park T and Casella G (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 103, 681–686.
- Polson NG, Scott JG, and Windle J (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association* 108, 1339–1349.

- Rockova V and George EI (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association* 109, 828–846.
- Rockova V and Lesaffre E (2014). Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics. *Bayesian Analysis* 9, 221–258.
- Sirvent A, Vigy O, Orsetti B, Urbach S, and Roche S (2012). Analysis of SRC oncogenic signaling in colorectal cancer by Stable Isotope Labeling with heavy Amino acids in mouse Xenografts. *Molecular & Cellular Proteomics* 11, 1937–1950. [PubMed: 23023324]
- Stingo FC, Chen YA, Tadesse MG, and Vannucci M (2011). Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes. *Annals of Applied Statistics* 5, 1978–2002. [PubMed: 23667412]
- Stingo FC and Vannucci M (2011). Variable Selection for Discriminant Analysis with Markov Random Field Priors for the Analysis of Microarray Data. *Bioinformatics* 27, 495–501. [PubMed: 21159623]
- Tang H, Wang Z, Liu Q, Liu X, Wu M, and Li G (2014). Disturbing miR-182 and -381 Inhibits BRD7 Transcription and Glioma Growth by Directly Targeting LRRC4. *PLoS ONE* 9,.
- Tibshirani R (1996). Regression Shrinkage and Selection via the Lasso. *Journal of Royal Statistical Society* 58, 267–288.
- Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *pdgfra*, *idh1*, *egfr*, and *nf1*. *Cancer cell* 17, 98–110. [PubMed: 20129251]
- Wu TT and Lange K (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* pages 224–244.
- Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society B* 68, 49–67.
- Zeng Y and Breheny P (2016). Overlapping Group Logistic Regression with Applications to Genetic Pathway Selection. *Cancer Informatics* 15, 179–187. [PubMed: 27679461]
- Zou H (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101, 1418–1429.

Figure 1.

Top two panels plot the marginal prior densities of β for (a) different μ while ν and σ are fixed and (b) different ν while μ and σ are fixed. Bottom two panels (c) and (d) plot the corresponding negative log density functions. The standard normal prior and the horseshoe prior with $\tau = 1$ are shown for contrast. The Laplacian prior with $\lambda = e^{0.3}$ is plotted as a comparison to the case with $\mu = 0.3$ and $\nu = 0.1$.

Figure 2. Contour plots of the marginal prior density of α_1 and α_2 for 4 different combinations of a_ω and b_ω .

Figure 3. Ranges of average values of estimated ω 's between important variables (I-I), between an important variable and an unimportant variable (I-U), and between unimportant variables (U-U) in EMSHS for scenario 1 and $p = 10,000$ case.

Table 1

The mean squared prediction error (MSPE) for the test data, the average false positive rate (FP), the average false negative rate (FN), both in percentage, and the average computation time per tuning parameter value in seconds are recorded. The methods in column 1 correspond to the Lasso, adaptive lasso, group lasso, the penalized approach by Li and Li (2008), the Markov random field approach for Bayesian variable selection incorporating structure knowledge, the EM approach by Rockova and George (2014) without and with structural knowledge, and the proposed methods. In the parentheses are the standard errors of the average differences in MSE from EMISH (Sec. 3 and 4) or EMSHS (Sec. 1, 2, and 5).

Scenario	Method	$p = 1,000$				$p = 10,000$				$p = 100,000$			
		MSPE	FP(%)	FN(%)	Time	MSPE	FP(%)	FN(%)	Time	MSPE	FP(%)	FN(%)	Time
1	Lasso	2.858	0.018	7.086	0.403	0.0	2.984(0.020)	2.143	0.142	0.1	3.009(0.021)	0.294	0.104
	ALasso	2.510	0.016	0.512	1.953	0.0	2.857(0.033)	0.105	3.134	0.2	3.680(0.064)	0.016	6.602
	GLasso	3.640	0.025	-	-	3.5	3.400(0.023)	-	-	102.3	3.438(0.026)	-	-
	Net	3.033	0.023	2.710	0.514	1.3	3.620(0.042)	0.419	1.508	120.4	-	-	444.0
	BVS-MRF	12.280	0.412	6.890	27.240	507.5	-	-	-	-	-	-	-
	EMVS	2.475	0.022	0.003	11.781	3.7	5.219(0.211)	0.000	38.096	112.6	-	-	-
	EMVSS	2.194	0.007	0.011	1.499	30.6	2.249(0.011)	0.003	3.289	701.7	-	-	-
	EMSH	2.257	0.011	0.021	3.399	0.1	2.376(0.019)	0.004	6.835	0.6	2.834(0.050)	0.001	13.452
	EMSHS	2.182	0.000	0.017	0.614	0.4	2.172(0.000)	0.002	1.215	2.9	2.200(0.000)	0.000	2.455
	Lasso	2.608	0.016	3.144	0.200	0.0	2.770(0.017)	1.223	0.226	0.1	2.871(0.019)	0.242	0.294
2	ALasso	2.273	0.008	0.274	0.357	0.0	2.357(0.013)	0.056	0.340	0.2	2.586(0.021)	0.008	1.042
	GLasso	3.640	0.027	-	-	3.4	3.422(0.026)	-	-	103.6	3.476(0.025)	-	456.4
	Net	2.775	0.024	1.908	0.215	1.2	3.024(0.027)	0.305	0.097	101.8	-	-	-
	BVS-MRF	7.647	0.269	9.057	16.819	508.8	-	-	-	-	-	-	-
	EMVS	2.204	0.007	0.006	2.224	2.9	2.670(0.051)	0.000	12.389	104.1	-	-	-
	EMVSS	2.180	0.004	0.019	0.440	29.0	2.162(0.008)	0.002	1.287	750.6	-	-	-
	EMSH	2.185	0.005	0.011	1.284	0.1	2.185(0.008)	0.001	3.192	0.6	2.257(0.012)	0.000	4.343
	EMSHS	2.158	0.000	0.002	0.307	0.4	2.122(0.000)	0.001	0.202	2.8	2.136(0.000)	0.000	0.323
	Lasso	2.858	0.019	7.086	0.403	0.0	2.984(0.024)	2.143	0.142	0.1	3.009(0.052)	0.294	0.104
	ALasso	2.510	0.013	0.512	1.953	0.0	2.857(0.027)	0.105	3.134	0.2	3.680(0.040)	0.016	6.602
3	GLasso	3.640	0.027	-	-	3.5	3.400(0.029)	-	-	102.3	3.438(0.055)	-	444.0
	Net	3.032	0.022	2.673	0.571	1.0	3.618(0.035)	0.416	1.635	119.7	-	-	-
	BVS-MRF	12.960	0.506	6.289	28.094	501.3	-	-	-	-	-	-	-

Scenario	Method	$p = 1,000$					$p = 10,000$					$p = 100,000$					
		MSPE	FP(%)	FN(%)	Time	MSPE	FP(%)	FN(%)	Time	MSPE	FP(%)	FN(%)	Time	MSPE	FP(%)	FN(%)	Time
4	EM VS	2.475	0.020	0.003	11.781	3.7	5.219(0.209)	0.000	38.096	112.6	-	-	-	-	-	-	-
	EMVSS	2.345	0.014	0.033	6.055	27.5	2.530(0.019)	0.016	8.565	726.9	-	-	-	-	-	-	-
	EMSH	2.257	0.000	0.021	3.399	0.1	2.376(0.000)	0.004	6.835	0.6	2.834(0.000)	0.001	13.452	8.4	0.001	13.452	8.4
	EMSHS	2.267	0.006	0.030	3.253	0.4	2.417(0.013)	0.007	6.723	2.5	2.873(0.019)	0.002	12.688	7.6	0.002	12.688	7.6
	Lasso	2.608	0.016	3.144	0.200	0.0	2.770(0.018)	1.223	0.226	0.1	2.871(0.020)	0.242	0.294	0.7	0.242	0.294	0.7
	ALasso	2.273	0.007	0.274	0.357	0.0	2.357(0.012)	0.056	0.340	0.2	2.586(0.017)	0.008	1.042	3.7	0.008	1.042	3.7
	GLasso	3.640	0.027	-	-	3.4	3.422(0.027)	-	-	103.6	3.476(0.027)	-	-	456.4	-	-	456.4
	Net	2.820	0.023	1.751	0.327	0.8	3.096(0.028)	0.286	0.444	115.4	-	-	-	-	-	-	-
	BVS-MRF	7.780	0.238	8.854	17.252	519.0	-	-	-	-	-	-	-	-	-	-	-
	EM VS	2.204	0.007	0.006	2.224	2.9	2.670(0.050)	0.000	12.389	104.1	-	-	-	-	-	-	-
5	EMVSS	2.209	0.007	0.022	1.573	26.6	2.268(0.013)	0.007	3.730	741.8	-	-	-	-	-	-	-
	EMSH	2.185	0.000	0.011	1.284	0.1	2.185(0.000)	0.001	3.192	0.6	2.257(0.000)	0.000	4.343	7.7	0.000	4.343	7.7
	EMSHS	2.215	0.006	0.016	2.145	0.4	2.191(0.005)	0.002	3.090	1.6	2.269(0.005)	0.000	4.479	9.3	0.000	4.479	9.3
	Lasso	2.858	0.018	7.086	0.403	0.0	2.984(0.020)	2.143	0.142	0.1	3.009(0.021)	0.294	0.104	0.5	0.294	0.104	0.5
	ALasso	2.510	0.016	0.512	1.953	0.0	2.857(0.033)	0.105	3.134	0.2	3.680(0.065)	0.016	6.602	3.6	0.016	6.602	3.6
	GLasso	3.640	0.025	-	-	3.5	3.400(0.024)	-	-	102.3	3.438(0.026)	-	-	444.0	-	-	444.0
	Net	3.032	0.023	2.707	0.514	0.6	3.620(0.042)	0.420	1.468	109.4	-	-	-	-	-	-	-
	BVS-MRF	12.426	0.465	6.854	29.054	506.6	-	-	-	-	-	-	-	-	-	-	-
	EM VS	2.475	0.022	0.003	11.781	3.7	5.219(0.211)	0.000	38.096	112.6	-	-	-	-	-	-	-
	EMVSS	2.192	0.006	0.008	1.499	28.5	2.249(0.011)	0.003	3.289	722.8	-	-	-	-	-	-	-
	EMSH	2.257	0.011	0.021	3.399	0.1	2.376(0.019)	0.004	6.835	0.6	2.834(0.051)	0.001	13.452	8.4	0.001	13.452	8.4
	EMSHS	2.174	0.000	0.007	0.599	0.4	2.164(0.000)	0.001	1.051	2.9	2.173(0.000)	0.000	1.637	13.1	0.000	1.637	13.1

Table 2

Cross-validated mean squared prediction error (CV MSPE) and computation time in seconds per tuning parameter (Time) from the analysis of TCGA genomic data and KEGG pathway information.

Method	CV MSPE	Time
Lasso	0.982	0.5
ALasso	1.066	3.9
GLasso	0.984	989.3
Net	0.985	115.6
EMVS	0.996	1346.6
EMVSS	0.984	5998.3
EMSH	0.968	7.0
EMSHS	0.970	35.9