Identification of Gene-environment interactions using a marginal robust Bayesian method

Xi Lu[1], **Kun Fan**[1] **and Cen Wu**[1,2]
[1] Department of Statistics, Kansas State University, Manhattan, KS
[2] School of Public Health, Yale University, 60 College ST, New Haven, CT 06520

Abstract

In high-throughput cancer studies, an important aim is to identify gene-environment interactions associated with the clinical outcomes. Recently, multiple marginal penalization methods have been developed and shown to be effective in G×E studies. However, within the Bayesian framework, marginal variable selection has not received much attention. In this study, we propose a novel marginal Bayesian variable selection method for G×E studies. In particular, our marginal Bayesian method is robust to data contamination and outliers in the outcome variables. With the incorporation of spike-and-slab priors, the proposed method outperforms a number of alternatives in both identification and prediction in extensive simulation studies. The utility of the marginal robust Bayesian variable selection method has been further demonstrated in the case studies using TCGA data. Some of the identified main and interaction effects from the real data analysis have important biological implications.

Keywords: Gene-environment interaction; marginal analysis; robust Bayesian variable selection; spike-and-slab priors.

1 Introduction

In high-throughput profiling studies, the significance of gene–environment (G×E) interactions in elucidating the genetic basis of complex diseases has been increasingly recognized beyond the main effects. In traditional studies, with Bayesian Lasso method, the shrinkage on the individual level of coefficients can be achieved. Although this method comes with many benefits, the main disadvantage is that they cannot shrink the posterior coefficient estimates to zero exactly. To overcome this difficult, we propose a robust Bayesian approach for variable selections in marginal model. We also incorporate spike-and-slab priors to impose sparsity. The advantages of the proposed method and model performance are evaluated through simulation. In the case study, the proposed Bayesian method is expected to lead to improve prediction and the identification of main and interaction effects with important implications. To facilitate fast computation and reproducible research, we implement the proposed and all alternative methods in C++ for the R package.

2 Data and Model Settings

We use Y to denote a continuous response variable representing the the cancer outcome or disease phenotype. Let $X = (X_1, \ldots, X_p)$ be the p genetic variants, $E = (E_1, \ldots, E_q)$ be the q environmental factors and $C = (C_1, \ldots, C_m)$ be the m clinical factors. We denote the ith subject with i. Let (Y_i, E_i, C_i, X_i) $(i = 1, \ldots, n)$ be independent and identically distributed random vectors. For the jth gene X_j $(j = 1, \ldots, p)$, define $W_j = (X_jE_1, \ldots, X_jE_q)$, $\eta_j = (\eta_{j1}, \ldots, \eta_{jq})^T$. Consider the following marginal model:

$$\begin{aligned} Y_i &= \sum_{k=1}^{q} \alpha_k E_{ik} + \sum_{t=1}^{m} \gamma_t C_{it} + \beta_j X_{ij} + \sum_{k=1}^{q} \eta_{jk} X_{ij} E_{ik} + \epsilon_i \\ &= \sum_{k=1}^{q} \alpha_k E_{ik} + \sum_{t=1}^{m} \gamma_t C_{it} + \beta_j X_{ij} + \eta_j W_j + \epsilon_i \end{aligned} \tag{1}$$

where α_k's and γ_t's are the regression coefficients corresponding to effects of environmental and clinical factors, respectively, and β_j's and η_{jk}'s are the regression coefficients of the genetic variants and G×E interactions effects, correspondingly. Denote $\alpha = (\alpha_1, \ldots, \alpha_q)^T$, $\gamma = (\gamma_1, \ldots, \gamma_m)^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$, $\eta = (\eta_1^T, \ldots, \eta_p^T)^T$, $W = (W_1, \ldots, W_p)$. Then model (1) can be written as

$$Y_i = E_i\alpha + C_i\gamma + X_{ij}\beta_j + W_j\eta_j + \epsilon_i. \tag{2}$$

2.1 Bayesian Robust method

The least absolute deviation (LAD) regression is well known for its advantages in dealing with long tailed distributions. Here, we propose a robust Bayesian method for variable selections in our marginal model. The Laplace distrubution in Bayesian LAD regression can be treated as a special case of the Laplace distribution in Baysian quantile regression. In Baysian quantile regression, we assume that $\epsilon_i (i = 1, \ldots, n)$ are i.i.d. random variables following the skewed Laplace distribution with density

$$f(\epsilon|\tau) = \theta(1-\theta)\tau exp(-\tau\rho_\theta(\epsilon))$$

The random errors can be written as

$$\epsilon_i = \xi_1 v_i + \tau^{-1/2}\xi_2\sqrt{v_i}z_i,$$

where

$$\xi_1 = \frac{1-2\theta}{\theta(1-\theta)} \quad and \quad \xi_1 = \sqrt{\frac{2}{\theta(1-\theta)}}$$

With $\theta \in (0, 1)$, $v_i \sim \exp(\tau^{-1})$, and $z_i \sim \mathrm{N}(0, 1)$.

As the Bayesian LAD regression is a special case of Bayesian quantile regression with θ=0.5, so we have $\xi_1 = 0$ and $\xi_2 = \sqrt{8}$. Therefore, the response Y_i can be written as:

$$\begin{aligned} Y_i &= \mu_i + \tau^{-1/2}\xi_2\sqrt{v_i}z_i, \\ v_i|\tau &\overset{iid}{\sim} \tau\exp(-\tau v_i), \\ z_i &\overset{iid}{\sim} \mathrm{N}(0, 1). \end{aligned} \tag{3}$$

Where $\mu_i = E_i\alpha + C_i\gamma + X_{ij}\beta_j + W_j\eta_j$.

2.2 Bayesian sparse variable selection and priors

In model (1), the coefficients β_j and η_j corresponds to the main and interaction effects with respect to the jth genentic variant, respectively. When $\beta_j = 0$ and $\eta_j = 0$, the genetic variant has no effect on the phenotype. A non-zero β suggests a presence of main effect while a non-zero η suggests an interaction effect. As the traditional method cannot shrink the posterior coefficient estimates to zero exactly, we incorporate spike-and-slab priors to impose sparsity.

For the robust Bayesian marginal model of the jth gene ($j = 1, \ldots, p$), consider the following priors:

$$\begin{aligned} \beta_j|s_1, \pi_1 &\sim (1-\pi_1)\mathrm{N}(0, s_1) + \pi_1\delta_0(\beta_j) \\ s_1|\varphi_1^2 &\sim \frac{\varphi_1^2}{2}\exp(-\frac{\varphi_1^2}{2}s_1) \\ \eta_{jk}|s_{2k}, \pi_2 &\overset{iid}{\sim} (1-\pi_2)\mathrm{N}(0, s_{2k}) + \pi_2\delta_0(\eta_k), (k = 1, \ldots, q) \\ s_{2k}|\varphi_2^2 &\overset{iid}{\sim} \frac{\varphi_2^2}{2}\exp(-\frac{\varphi_2^2}{2}s_{2k}), (k = 1, \ldots, q) \end{aligned} \tag{4}$$

Here, π_1 and π_2 control the sparisity on the main and interaction level, repectively. The prior can be non-informative if π_1 and π_2 are given values with 0.5 as their priors are given the same probability. So we assign $\pi_1 \sim \mathrm{Beta}(r_1, u_1)$ and $\pi_2 \sim \mathrm{Beta}(r_2, u_2)$ with these conjugate beta priors which account for the uncertainty in π_1 and π_2. In this paper, we choose $r_1 = u_1 = r_2 = u_2 = 1$.

We place normal priors on $\alpha_k (k = 1, \ldots, q)$ and $\gamma_t (t = 1, \ldots, m)$ as

$$\begin{aligned} \alpha_k &\overset{iid}{\sim} \frac{1}{\sqrt{(2\pi\alpha_0)}}\exp(-\frac{\alpha_k^2}{2\alpha_0}), (k = 1, \ldots, q) \\ \gamma_t &\overset{iid}{\sim} \frac{1}{\sqrt{(2\pi\gamma_0)}}\exp(-\frac{\gamma_t^2}{2\gamma_0}), (t = 1, \ldots, m) \end{aligned}$$

We also assume Gamma priors on τ, φ_1^2 and φ_2^2 with

$$\begin{aligned} \tau &\sim \mathrm{Gamma}(a, b), \\ \varphi_1^2 &\sim \mathrm{Gamma}(c_1, d_1), \\ \varphi_2^2 &\sim \mathrm{Gamma}(c_2, d_2). \end{aligned}$$

2.3 Computation

Denote $\tilde{W} = W_j$, for the jth gene, the joint posterior distribution of all the unknown parameters conditional on data can be expressed as

$$
\begin{aligned}
&\pi(\alpha, \gamma, \beta_j, \eta_j, v, s_1, s_2, \tau, \varphi_1, \varphi_2, \pi_1, \pi_2, z_i|Y) \\
&\qquad \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2 v_i}} \exp\Big\{ - \frac{(y_i - E_i\alpha - C_i\gamma - X_{ij}\beta_j - \tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2 v_i} \Big\} \\
&\qquad \times \prod_{i=1}^{n} \tau \exp(-\tau v_i) \tau^{a-1} \exp(-b\tau) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z_i^2) \\
&\qquad \times \prod_{k=1}^{q} \frac{1}{\sqrt{(2\pi\alpha_0)}} \exp(-\frac{\alpha_k^2}{2\alpha_0}) \\
&\qquad \times \prod_{t=1}^{m} \frac{1}{\sqrt{(2\pi\gamma_0)}} \exp(-\frac{\gamma_t^2}{2\gamma_0}) \\
&\qquad \times \Big((1-\pi_1)(2\pi s_1)^{-1/2} \exp(-\frac{\beta_j^2}{2s_1}) \mathbf{I}_{\{\beta_j \neq 0\}} + \pi_1 \delta_0(\beta_j) \Big) \\
&\qquad \times \prod_{k=1}^{q} \Big((1-\pi_2)(2\pi s_{2k})^{-1/2} \exp(-\frac{\eta_{jk}^2}{2s_{2k}}) \mathbf{I}_{\{\eta_{jk} \neq 0\}} + \pi_2 \delta_0(\eta_{jk}) \Big) \\
&\qquad \times \frac{\varphi_1^2}{2} \exp(-\frac{\varphi_1^2}{2} s_1) \\
&\qquad \times \prod_{k=1}^{q} \frac{\varphi_2^2}{2} \exp(-\frac{\varphi_2^2}{2} s_{2k}) \\
&\qquad \times (\varphi_1^2)^{c_1 - 1} \exp(-d_1 \varphi_1^2) \\
&\qquad \times (\varphi_2^2)^{c_2 - 1} \exp(-d_2 \varphi_2^2) \\
&\qquad \times \pi_1^{r_1 - 1} (1-\pi_1)^{u_1 - 1} \\
&\qquad \times \pi_2^{r_2 - 1} (1-\pi_2)^{u_2 - 1}
\end{aligned}
$$

Let $\mu_{(-\alpha_k)} = E(Y_i) - E_{ik}\alpha_k, (i = 1, \dots, n), (k = 1, \dots, q)$, representing the mean effect without the contribution of $E_{ik}\alpha_k$. The posterior distribution of α_k conditional on all other parmeters can be expressed as

$$
\begin{aligned}
&\pi(\alpha_k|\text{rest}) \\
&\quad \propto \pi(\alpha_k)\pi(y|\cdot) \\
&\quad \propto \exp\Big\{ - \sum_{i=1}^{n} \frac{(y_i - E_i\alpha - C_i\gamma - X_{ij}\beta_j - \tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2 v_i} \Big\} \times \exp(-\frac{\alpha_k^2}{\alpha_0}) \\
&\quad \propto \exp\Big\{ -\frac{1}{2} [(\sum_{i=1}^{n} \frac{\tau E_{ik}^2}{\xi_2^2 v_i} + \frac{1}{\alpha_0})\alpha_k^2 - 2\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\alpha_k)})E_{ik}}{\xi_2^2 v_i} \alpha_k] \Big\}
\end{aligned}
$$

Hence, the full conditional distribution of α_k is normal distribution $N(\mu_{\alpha_k}, \sigma^2_{\alpha_k})$with mean

$$\mu_{\alpha_k} = (\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\alpha_k)})E_{ik}}{\xi_2^2 v_i})\sigma^2_{\alpha_k},$$

and variance

$$\sigma^2_{\alpha_k} = (\sum_{i=1}^{n} \frac{\tau E_{ik}^2}{\xi_2^2 v_i} + \frac{1}{\alpha_0})^{-1}.$$

The posterior distribution of $\gamma_t (t = 1, \dots, m)$ conditional on all other parameters can be obtained in similiar way.

$$\gamma_t|\text{rest}{\sim}N(\mu_{\gamma_k}, \sigma^2_{\gamma_t})$$

where

$$\mu_{\gamma_t} = (\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\gamma_t)})C_{it}}{\xi_2^2 v_i})\sigma^2_{\gamma_t},$$

$$\sigma^2_{\gamma_t} = (\sum_{i=1}^{n} \frac{\tau C_{it}^2}{\xi_2^2 v_i} + \frac{1}{\gamma_0})^{-1}.$$

Let $\mu_{(-\beta_j)} = E(Y_i) - X_{ij}\beta_j$ and $l_1 = \pi(\beta_j = 0|\text{rest})$, the conditional posterior distribution of β_j is a multivariate spike-and-slab distribution:

$$\beta_j|\text{rest}{\sim}(1 - l_1)N(\mu_{\beta_j}, \sigma^2_{\beta_j}) + l_1\delta_0(\beta_j) \tag{5}$$

where

$$\mu_{\beta_j} = (\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\beta_j)})X_{ij}}{\xi_2^2 v_i})\sigma^2_{\beta_j},$$

$$\sigma^2_{\beta_j} = (\sum_{i=1}^{n} \frac{\tau X_{ij}^2}{\xi_2^2 v_i} + \frac{1}{s_1})^{-1}.$$

It's easy to show that

$$l_1 = \frac{\pi_1}{\pi_1 + (1 - \pi_1)s_1^{-1/2}(\sigma^2_{\beta_j})^{1/2}\text{exp}\{\frac{1}{2}(\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\beta_j)})X_{ij}}{\xi_2^2 v_i})^2\sigma^2_{\beta_j}\}}.$$

The posterior distribution of β_j is a mixture of a multivariate normal distribution and a point mass at 0. That is, at each iteractio of MCMC, β_j is drawn from $N(\mu_{\beta_j}, \sigma^2_{\beta_j})$ with probability $(1 - l_1)$ and is set to 0 with probability l_1.

Similiarly, the posterior distribution of $\eta_{jk}(k = 1, \dots, q)$ is also a spike-and-slab distribution. Denote $\mu_{(-\eta_{jk})} = E(Y_i) - W_{ik}\eta_{jk}$ and $l_{2k} = \pi(\eta_{jk} = 0|\text{rest})$, η_{jk} follows this distribution:

$$\eta_{jk}|\text{rest}{\sim}(1 - l_{2k})N(\mu_{\eta_{jk}}, \sigma^2_{\eta_{jk}}) + l_{2k}\delta_0(\eta_{jk}) \tag{6}$$

where

$$\mu_{\eta_{jk}} = (\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\eta_{jk})})\tilde{W}_{ik}}{\xi_2^2 v_i})\sigma^2_{\eta_{jk}},$$

$$\sigma^2_{\beta_j} = (\sum_{i=1}^{n} \frac{\tau \tilde{W}_{ik}^2}{\xi_2^2 v_i} + \frac{1}{s_{2k}})^{-1}.$$

And

$$l_{2k} = \frac{\pi_2}{\pi_2 + (1-\pi_2)s_{2k}^{-1/2}(\sigma^2_{\eta_{jk}})^{1/2}\text{exp}\{\frac{1}{2}(\sum_{i=1}^{n}\frac{\tau(y_i-\mu_{(-\eta_{jk})})\tilde{W}_{ik}}{\xi_2^2 v_i})^2\sigma^2_{\eta_{jk}}\}}. \tag{7}$$

The full conditional posterior distribution of s_1 is:

$$\begin{aligned}
&s_1|\text{rest}\\
&\quad\propto \pi(\beta|s_1,\pi_1)\pi(s_1)\\
&\quad\propto \text{exp}(-\frac{\varphi_1^2}{2}s_1)\Big((1-\pi_1)(2\pi s_1)^{-1/2}\text{exp}(-\frac{\beta_j^2}{2s_1})\mathbf{I}_{\{\beta_j\neq 0\}} + \pi_1\delta_0(\beta_j)\Big)
\end{aligned} \tag{8}$$

When $\beta_j = 0$, (8) is proportion to $\text{exp}(-\frac{\varphi_1^2}{2}s_1)$. Therefore, the posterior distribution of s_1 is $\text{exp}(\frac{\varphi_1^2}{2})$.

When $\beta_j \neq 0$, (8) is proportion to

$$\begin{aligned}
&\frac{1}{\sqrt{s_1}}\text{exp}(-\frac{\varphi_1^2}{2}s_1)\text{exp}(-\frac{\beta_j^2}{2s_1})\\
&\qquad\propto \frac{1}{\sqrt{s_1}}\text{exp}\Big\{-\frac{1}{2}[\varphi_1^2 s_1 + \frac{\beta_j^2}{s_1}]\}
\end{aligned}$$

Therefore, when $\beta_j \neq 0$, the posterior distribution for s_1^{-1} is Inverse-Gaussian($\sqrt{\frac{\varphi_1^2}{\beta_j^2}}, \varphi_1^2$).

Similiarly, for $s_{2k}(k = 1, \ldots, q)$, when $\eta_{jk} = 0$, the posterior distribution of s_{2k} is $\text{exp}(\frac{\varphi_2^2}{2})$. When $\eta_{jk} \neq 0$, the posterior distribution for s_{2k}^{-1} is Inverse-Gaussian($\sqrt{\frac{\varphi_2^2}{\eta_{jk}^2}}, \varphi_2^2$).

The full conditional posterior distribution of φ_1^2:

$$\begin{aligned}
&\varphi_1^2|\text{rest}\\
&\quad\propto \pi(s_1|\varphi_1^2)\pi(\varphi_1^2)\\
&\quad\propto \frac{\varphi_1^2}{2}\text{exp}(-\frac{\varphi_1^2 s_1}{2})(\varphi_1^2)^{c_1-1}\text{exp}(-d_1\text{exp})\\
&\quad\propto (\varphi_1^2)^{c_1}\text{exp}\Big(-\varphi_1^2(s_1/2+d_1)\Big)
\end{aligned}$$

Therefore, the posterior distribution for φ_1^2 is Gamma($c_1+1, s_1/2+d_1$). Similiarly, the posterior distribution for φ_2^2 is Gamma($c_2+q, \sum_{k=1}^{q} s_{2k}/2+d_2$).

The full conditional posterior distribution of π_1:

$$\begin{aligned}
&\pi_1|\text{rest}\\
&\quad\propto \pi(s_1|\varphi_1^2)\pi(\varphi_1^2)\\
&\quad\propto \pi_1^{r_1-1}(1-\pi_1)^{u_1-1}\\
&\quad\times\Big((1-\pi_1)(2\pi s_1)^{-1/2}\text{exp}(-\frac{\beta_j^2}{2s_1})\mathbf{I}_{\{\beta_j\neq 0\}} + \pi_1\delta_0(\beta_j)\Big)
\end{aligned}$$

Then, the posterior distribution for π_1 is Beta $(1 + r_1 - \mathbf{I}(\beta_j \neq 0), u_1 + \mathbf{I}(\beta_j \neq 0))$.

The full conditional posterior distribution of π_2:

$$\begin{aligned}
&\pi_2|\text{rest}\\
&\quad\propto \pi(s_2|\varphi_2^2)\pi(\varphi_2^2)\\
&\quad\propto \pi_2^{r_2-1}(1-\pi_2)^{u_2-1}\\
&\quad\times\prod_{k=1}^{q}\Big((1-\pi_2)(2\pi s_{2k})^{-1/2}\exp(-\frac{\eta_{jk}^2}{2s_{2k}})\mathbf{I}_{\{\eta_{jk}\neq 0\}}+\pi_2\delta_0(\eta_{jk})\Big)
\end{aligned}$$

So, the posterior distribution for π_2 is Beta $(1+r_1-\sum_{k=1}^{q}\mathbf{I}(\eta_{jk}\neq 0), u_1+\sum_{k=1}^{q}\mathbf{I}(\eta_{jk}\neq 0))$.

The full conditional posterior distribution of τ:

$$\begin{aligned}
&\tau|\text{rest}\\
&\quad\propto \pi(v|\tau)\pi(\tau)\pi(y|\cdot)\\
&\quad\propto \tau^{n/2}\exp\Big\{-\sum_{i=1}^{n}\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2v_i}\Big\}\\
&\quad\times\tau^{n}\exp(-\tau\sum_{i=1}^{n}v_i)\tau^{a-1}\exp(-b\tau)\\
&\quad\propto \tau^{a+\frac{3}{2}n-1}\exp\Big\{-\tau\big[\sum_{i=1}^{n}(\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\xi_2^2v_i}+v_i)+b\big]\Big\}
\end{aligned}$$

Therefore, the posterior distribution for τ is Gamma$(a+\frac{3}{2}n,\ \big[\sum_{i=1}^{n}(\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\xi_2^2v_i}+v_i)+b\big])$.

Last, we have The full conditional posterior distribution of v_i:

$$\begin{aligned}
&v_i|\text{rest}\\
&\quad\propto \pi(v|\tau)\pi(y|\cdot)\\
&\quad\propto \frac{1}{\sqrt{v_i}}\exp\Big\{-\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2v_i}\Big\}\times\exp(-\tau v_i)\\
&\quad\propto \frac{1}{\sqrt{v_i}}\exp\Big\{-\frac{1}{2}\big[(2\tau)v_i+\frac{\tau(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{\xi_2^2v_i}\big]\Big\}
\end{aligned}$$

It is easy to show that

$$\frac{1}{v_i}\sim\text{ Inverse-Gaussian}(\sqrt{\frac{2\xi_2^2}{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}},\ 2\tau).$$

3 Simulation

To demonstrate the utility of the proposed approach, we evaluate the performance through simulation study. In particular, we compare the performance of the proposed method, robust

Bayesian Lasso spike-and-slab variable selection (denoted as LADBLSS) with three alternatives, robust Bayesian Lasso variable selection (denoted as LADBL), Bayesian Lasso with spike-and-slab variable selection (denoted as BLSS) and Bayesian Lasso variable selection (denoted as BL). LADBL is similar to the proposed method, except that it does not adopt the spike-and-slab prior. Comparison of LADBLSS with BLSS demonstrate the importance of accommodating outliers.

We consider two data settings in our simulation. In the first setting, we have continuous environmental factors with true non-zero signal of interactions close to each other. In the second setting, we have one half of the environmental factors which is continuous and another half is discrete with the true non-zero signal of interactions far away to each other. For each setting, we consider four environmental factors. For continuous E, we simulate normally distributed factors with mean 0 and variance 1. The correlation between the jth and kth continuous E factors is $\rho^{|j-k|}$ with $\rho = 0.5$. In addition, we simulate $m = 3$ clinical factors from a multivariate normal distribution with margianl mean 0 and marginal variance 1 and AR (auto-regressive) correlation structure with $\rho = 0.5$. In simulating continuous genetic variants, we generate multivariate normal distributions with marginal mean 0 and variance 1. The AR structure is considered in computing the correlation of G factors, under which gene j and k have correlation $\rho^{|j-k|}$ with $\rho = 0.5$. In these setting, the sample size is set as $n = 200$, and the number of G factors $p = 500$ and $p = 1000$ are both considered. Among the p main G effects and $p \times q$ $G \times E$ interactions, 8 and 12 are set as associated with the response, respectively. All environmental factors have important main effects. The nonzero coefficients of important effects are randomly generateed from a uniform distribution $Unif[0.1, 0.5]$. The random error are generated from: (1) $N(0,1)$(Error 1), (2) t-distribution with 2 degress of freedom ($t(2)$)(Error2), (3) LogNormal(0,2)(Error3), (4) 90%$N(0,1)$+10%Cauchy(0,1)(Error4), (5) 80%$N(0,1)$+20%Cauchy(0,1)(Error5). All of them are heavy-tailed distribution except the first one.

Posterior samples are collected from a Gibbs Sampler with 10,000 interations in which the first 5,000 are burn-ins. The posterior medians are used to estimate the coefficients. For methods incorporating spike-and-slab priors, we consider the median probability model (MPM) to identify significant predictors. Here we use ϕ as the indicator identifying the probability of the posterior distribution of predictor generating from non-spike distribution. Suppose we collect G posterior samples from MCMC after burn-ins. The jth predictor is included in the regression model at gth MCMC iteraction if the indicator of this predictor at this step is 1, that is if $\phi_j^{(g)} = 1$. Then the probability of the posterior distribution including the jth predictor in the final model is defined as the average of all the indicators for the jth predictor among the G posterior samples. That is

$$p_j = \hat{\pi}(\phi_j = 1|y) = \frac{1}{G}\sum_{g=1}^{G} \phi_j^{(g)},\ j = 1, \ldots, p$$

A higher posterior inclusion probability p_j indicates a stronger empirical evidence that the jth predictor has a non-zero coefficient, which also indicates that this predictor has a stronger association with the response variable. Usually, the MPM is defined with the predictors which have posterior inclusion probabiliry no less than $\frac{1}{2}$. For methods without spike-and-slab priors, the 95% credible interval is used.

To compare the performance of the four approaches, we consider a sequence of cutting-off probabilities in MPM for methods with spike-and-slab priors and different credible intervals for methods without spike-and-slab priors, compute true-positive-rate (TPR) and false-positive-rate (FPR) values, and use the area under curve (AUC) under the receiver operating characteristic (ROC) framework to compare the identification accuracy. In addition, we also consider Top100, which is defined as the number of true signals when 100 important main effects (or interactions) are identified.

Table 1: Simulation results of the first setting. AUC (mean of AUC), SD (sd of AUC) based on 100 replicates, p=500 and n=200.

		BL	BLSS	LADBL	LADBLSS
Error 1	AUC	0.9182	0.9901	0.9258	0.9887
	SD	0.0052	0.0021	0.0076	0.0026
Error 2	AUC	0.8332	0.9420	0.9004	0.9841
	SD	0.0107	0.0235	0.0078	0.0031
Error 3	AUC	0.5343	0.5473	0.8432	0.9558
	SD	0.0144	0.0576	0.0115	0.0161
Error 4	AUC	0.8221	0.9124	0.9222	0.9895
	SD	0.0212	0.0410	0.0071	0.0024
Error 5	AUC	0.7507	0.8431	0.9192	0.9904
	SD	0.0217	0.0633	0.0059	0.0018

Table 2: Simulation results of the second setting. AUC (mean of AUC), SD (sd of AUC) based on 100 replicates, p=500 and n=200.

		BL	BLSS	LADBL	LADBLSS
Error 1	AUC	0.8413	0.8995	0.8294	0.8814
	SD	0.0066	0.0179	0.0096	0.0101
Error 2	AUC	0.7716	0.8138	0.8092	0.8598
	SD	0.0085	0.0288	0.0073	0.0123
Error 3	AUC	0.5385	0.4917	0.7654	0.8001
	SD	0.0123	0.0403	0.0127	0.0212
Error 4	AUC	0.7620	0.7679	0.8263	0.8715
	SD	0.0096	0.0635	0.0078	0.0141
Error 5	AUC	0.7121	0.6995	0.8201	0.8675
	SD	0.0167	0.0765	0.0088	0.0129

We can observe that the proposed model has better performance over the other three when dealing with heavy-tailed distributions.

Table 3: Identification results of the first setting with Top100 method. mean(sd) based on 100 replicates, p=500 and n=200.

		Main	Interaction	Total
Error 1	BL	7.6(0.49)	6.8(1.6)	14.4(1.73)
N(0,1)	BLSS	7.8(0.41)	10.8(0.92)	18.6(1.13)
	LADBL	7.67(0.55)	6.53(1.85)	14.2(1.81)
	LADBLSS	7.76(0.5)	10.53(1.36)	18.3(1.49)
Error 2	BL	6.37(1.90)	3.9(2.07)	10.27(3.19)
$t(2)$	BLSS	6.33(1.63)	8.53(2.46)	14.87(3.71)
	LADBL	7.43(0.94)	5.8(1.71)	13.23(2.01)
	LADBLSS	7.53(0.51)	9.9(1.56)	17.43(1.76)
Error 3	BL	0.9(1.21)	0.5(0.97)	1.4(1.45)
Lognormal(0,2)	BLSS	0.73(0.94)	0.47(0.68)	1.2(1.35)
	LADBL	6.27(1.55)	3.67(1.94)	9.93(2.75)
	LADBLSS	6.1(1.37)	8.93(2.02)	15.03(3.09)
Error 4	BL	5.57(2.99)	3.63(2.53)	9.2(5.05)
90%N(0,1)	BLSS	6.2(2.62)	8.3(3.98)	14.5(6.39)
+10%Cauchy(0,1)	LADBL	7.77(0.43)	7.00(1.93)	14.77(1.81)
	LADBLSS	7.77(0.57)	10.67(1.50)	18.23(1.67)
Error 5	BL	5.07(2.89)	3(2.49)	8.07(5.01)
80%N(0,1)	BLSS	4.6(3.25)	5.7(4.23)	10.3(7.27)
+20%Cauchy(0,1)	LADBL	7.57(0.57)	6.83(1.07)	14.4(1.83)
	LADBLSS	7.8(0.55)	10.53(1.36)	18.33(1.69)

Table 4: Identification results of the second setting with Top100 method. mean(sd) based on 100 replicates, p=500 and n=200.

		Main	Interaction	Total
Error 1	BL	7.83(0.46)	3.80(1.09)	11.63(1.10)
N(0,1)	BLSS	7.90(0.31)	5.07(1.41)	12.97(1.35)
	LADBL	7.90(0.40)	3.60(1.49)	11.50(1.63)
	LADBLSS	7.33(0.52)	4.03(1.59)	11.76(1.48)
Error 2	BL	6.90(1.88)	1.87(1.48)	8.77(2.75)
$t(2)$	BLSS	6.97(1.27)	3.00(1.74)	9.97(2.35)
	LADBL			
	LADBLSS	7.47(0.82)	3.9(1.29)	11.37(1.59)
Error 3	BL	0.53(0.89)	0.53(0.73)	1.07(1.08)
Lognormal(0,2)	BLSS	0.63(0.81)	0.77(1.00)	1.40(1.19)
	LADBL	7.10(1.03)	1.60(1.25)	8.70(1.66)
	LADBLSS	6.73(1.20)	2.97(1.13)	9.70(1.95)
Error 4	BL	6.73(2.48)	1.87(1.36)	8.60(3.45)
90%N(0,1)	BLSS	6.10(2.86)	3.00(2.12)	9.10(4.51)
+10%Cauchy(0,1)	LADBL	7.97(0.18)	3.17(1.78)	11.13(1.74)
	LADBLSS	7.77(0.57)	4.53(1.69)	12.30(1.84)
Error 5	BL	6.13(2.33)	2.27(1.34)	8.40(2.91)
80%N(0,1)	BLSS	4.00(3.18)	1.43(1.50)	5.43(4.46)
+20%Cauchy(0,1)	LADBL	7.8(0.41)	2.73(1.44)	10.53(1.50)
	LADBLSS	7.80(0.48)	4.37(1.38)	12.17(1.51)

Compared with the alternative approaches, it can observe that the proposed model has higher average of TP (true-positive) over the other three when we have heavy-tailed distributions.

4 Real Data Analysis

In this study, we consider skin cutaneous melanoma (SKCM) from the Cancer Genome Atlas (TCGA), which is organized by the National Cancer Institute (NCI) with high quality genetic, clinical and proteomic data. We use the level-3 gene expression data of SKCM from the cBio Cancer Genomics Portal. Messenger RNA (mRNA) gene expressions are used as G factor. For E factor, we consider Age, AJCC pathologic tumor stage, gender and Clark level. The response variable is the log-transformed Breslow's thickness. Data are available on 298 subjects and 18,934 gene expressions among which 10,000 genes with the strongest association with the response variables are selected for $G \times E$ interaction analysis. We are trying to identify important gene expressions that have significant main effect or $G \times E$ interaction effects on the Breslow's thinkness.

Table 5: The numbers of main G effects and interactions identified by different approaches and their overlaps.

SKCM	Main				Interaction			
	BL	BLSS	LADBL	LADBLSS	BL	BLSS	LADBL	LADBLSS
BL								
BLSS								
LADBL								
LADBLSS								

Table 6: Analysis of SKCM with proposed method: identified main and interaction effects

		Interactions			
Gene	Main Effects	Clark level	AJCC stage	Age	Gender
AKR1C1					
CPS1					
PSPH					
IGHD					
IFI27					

5 Discussion

References

[1] Wu, C. and Ma, S. (2014) A selective review of robust variable selection with applications in bioinformatics. *Brief. Bioinform* 1–11. doi: 10.1093/bib/bbu046

[2] Huang J., Ma S. and Xie H. (2007). Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*, **17**: 1533–1548.

[3] Wu et. al. (2015) A robust network–constrained penalization approach for integrative analysis with applications in TCGA data. (Submitted)

A Appendix

A.1 The ROC curves in simulation

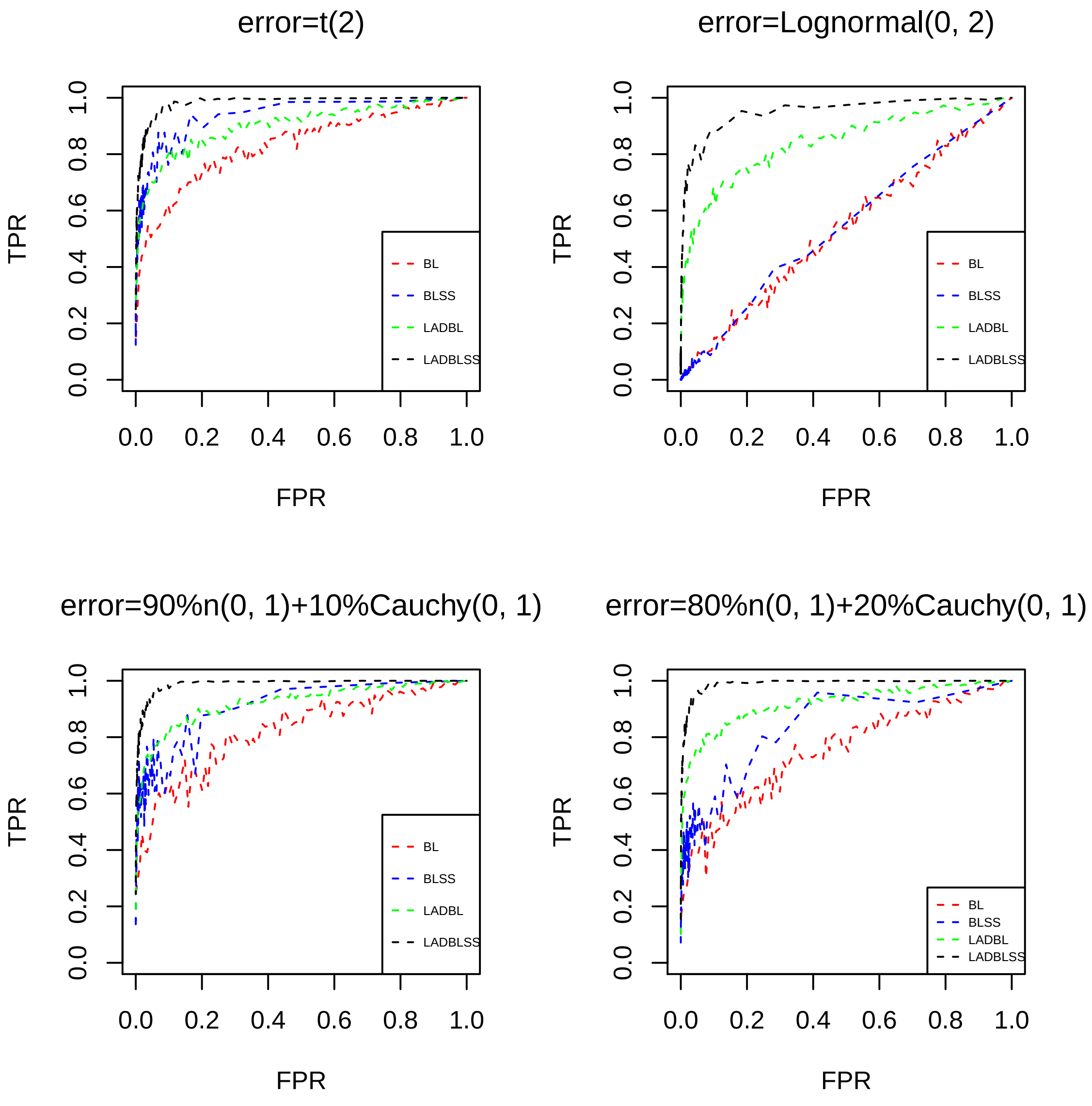

Figure 1: ROC curves of the first setting in simulation

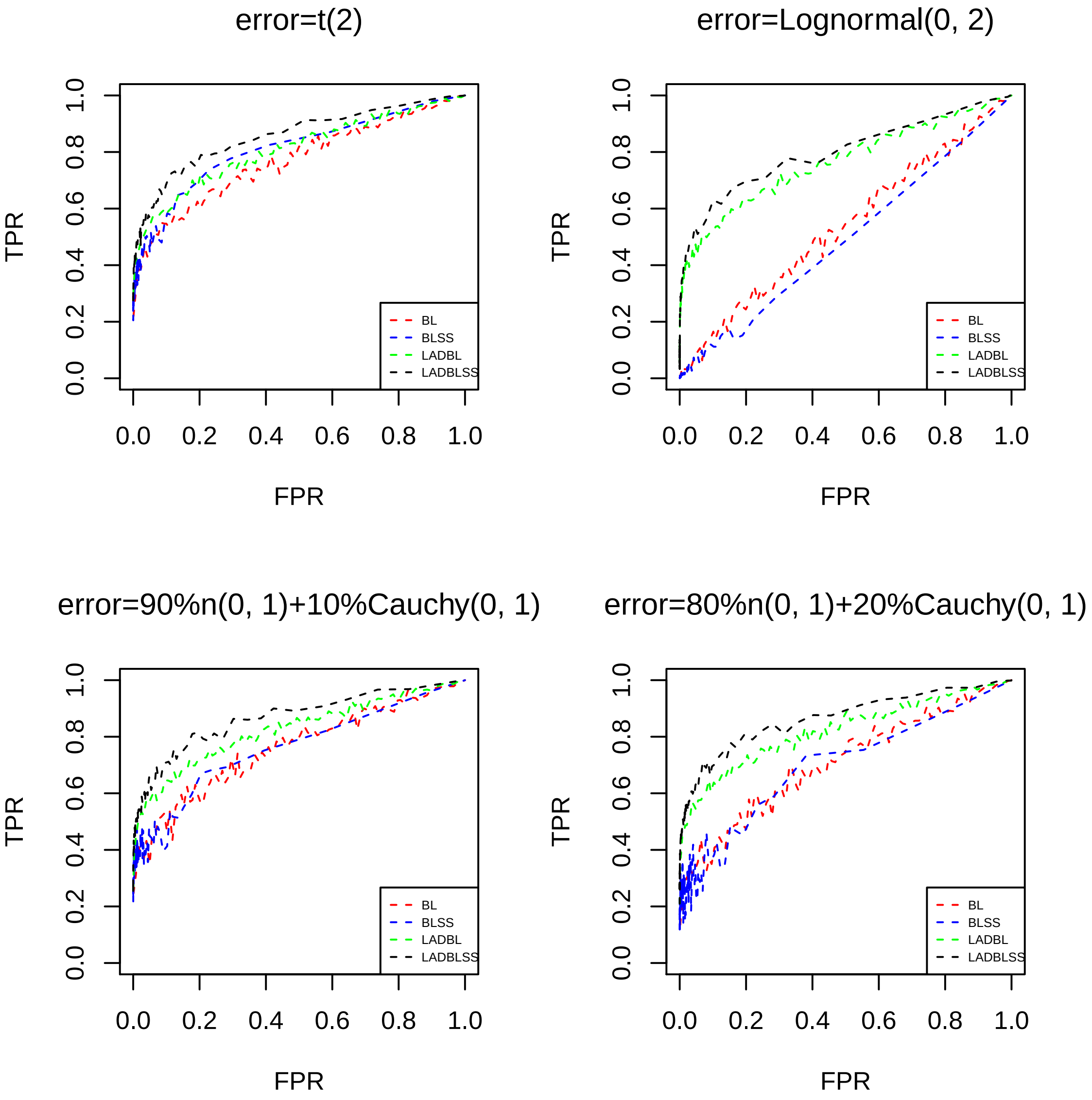

Figure 2: ROC curves of the second setting in simulation