HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2024 June 01.

Published in final edited form as:

Biometrics. 2023 June ; 79(2): 684–694. doi:10.1111/biom.13670.

Robust Bayesian variable selection for gene–environment interactions

Jie Ren[1], **Fei Zhou**[2], **Xiaoxi Li**[2], **Shuangge Ma**[3], **Yu Jiang**[4], **Cen Wu**[2]

[1]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, Indiana, USA

[2]Department of Statistics, Kansas State University, Manhattan, Kansas, USA

[3]Department of Biostatistics, Yale University, New Haven, Connecticut, USA

[4]Division of Epidemiology, Biostatistics and Environmental Health, School of Public Health, University of Memphis, Memphis, Tennessee, USA

Abstract

Gene–environment (G×E) interactions have important implications to elucidate the etiology of complex diseases beyond the main genetic and environmental effects. Outliers and data contamination in disease phenotypes of G×E studies have been commonly encountered, leading to the development of a broad spectrum of robust regularization methods. Nevertheless, within the Bayesian framework, the issue has not been taken care of in existing studies. We develop a fully Bayesian robust variable selection method for G×E interaction studies. The proposed Bayesian method can effectively accommodate heavy-tailed errors and outliers in the response variable while conducting variable selection by accounting for structural sparsity. In particular, for the robust sparse group selection, the spike-and-slab priors have been imposed on both individual and group levels to identify important main and interaction effects robustly. An efficient Gibbs sampler has been developed to facilitate fast computation. Extensive simulation studies, analysis of diabetes data with single-nucleotide polymorphism measurements from the Nurses' Health Study, and The Cancer Genome Atlas melanoma data with gene expression measurements demonstrate the superior performance of the proposed method over multiple competing alternatives.

Keywords

Bayesian variable selection; gene–environment interactions; Markov chain Monte Carlo; robust analysis; sparse group selection

Correspondence Cen Wu, Department of Statistics, Kansas State University, Manhattan, KS 66506, USA. wucen@ksu.edu.

SUPPORTING INFORMATION

Web Appendices A, B, C and D, Tables and Figures referenced in Sections 2–4, are available with this paper at the Biometrics website on Wiley Online Library. The R package "roben" that implements all the methods under comparison is available both on CRAN (http://cran.r-project.org/package=roben) and Wiley Online Library.

1 | INTRODUCTION

Deciphering the genetic architecture of complex diseases is a challenging task, as it demands the elucidation of the coordinated function of multiple genetic factors, their interactions, as well as gene–environment (G×E) interactions. How the genetic contributions to influence the variations in the disease phenotypes are mediated by the environmental factors reveals a unique perspective of the disease etiology beyond the main genetic effects and their interactions (or epistasis) (Hunter, 2005; Simonds et al., 2016). Till now, G×E interaction analyses have been extensively conducted, especially within the framework of genetic association studies, to search for the important main and interaction effects that are associated with the disease trait (Mukherjee et al., 2011).

With the availability of a large number of genetic factors, such as single-nucleotide polymorphisms (SNPs) or gene expression, G×E interactions are of high dimensionality even though the preselected environmental factors are usually low dimensional. Therefore, variable selection has emerged as a powerful tool to identify G×E interactions associated with phenotypic traits (Fan and Lv, 2010; Wu and Ma, 2015), and an increasing amount of G×E studies have recently been conducted along this line, especially with regularization methods (Zhou et al., 2021).

A prominent trend among these studies is to incorporate robustness in regularized identification of main and interaction effects in order to accommodate data contamination and heavy-tailed distributions in the disease phenotypes. Using the datasets analyzed in this article as an example, the disease outcomes of interest are weight from the Nurses' Health Study (NHS) and (log-transformed) Breslow's depth from The Cancer Genome Atlas (TCGA) Skin Cutaneous Melanoma (SKCM) data. We generate the box-plots (Figure 1) and histograms (Web Figure 1) of residuals corresponding to the two disease phenotypes using Bayesian linear regression with clinical covariates and five genetic variates that have the most significant p-values in the marginal model. The long tails can be clearly observed. In practice, such a heavy-tailed distribution is frequently encountered and arises due to multiple reasons. For instance, some phenotypes have skewness in nature. For the subjects recruited for the NHS, their ages are in the range from 41 to 68 as the average age for the onset of type 2 diabetes is 45 (Centers for Disease Control and Prevention, 2020). The subjects' weight among this age group does have a right-skewed tendency. In addition, in the study of complex diseases such as cancer, even patients of similar profiles may have different subtypes as rigorous recruitment of patients is usually not possible on the grounds of cost. The data from the major disease subtype can be viewed as being "contaminated" by other subtypes or outliers. As nonrobust approaches cannot efficiently accommodate data contamination and long-tailed distributions, which inevitably leads to biased estimates and false identifications, the robust regularization methods have thus been extensively developed for G×E studies (Wu and Ma, 2015; Zhou et al., 2021).

Nevertheless, within the Bayesian framework, robust variable selection methods have not yet been investigated for gene–environment interactions. In fact, our literature search indicates that only limited number of Bayesian variable selection methods have been developed for G×E studies and none of them is robust (Zhou et al., 2021). Driven by the urgent

need to conduct robust Bayesian analysis, we propose robust Bayesian variable selection methods tailored for interaction studies by adopting a Bayesian formulation of the least absolute deviation (LAD) regression to accommodate data contamination and long-tailed distributions in the phenotype. Such a formulation is a special case of the Bayesian quantile regression (Yu and Moyeed, 2001). The LAD loss has been a very popular choice for developing robust regularization methods for data with structured sparsity, including networks (Wu et al., 2018; Ren et al., 2019) and sparse group structure (Wu et al., 2018). Its computational convenience has been revealed within the Bayesian framework as an efficient Gibbs sampler can be constructed when the loss is combined with LASSO, group LASSO, and elastic net penalties (Li et al., 2010). Furthermore, following the strategy of bi-level selection from a nonrobust Bayesian setting (Xu and Ghosh, 2015), we have developed the Bayesian LAD sparse group selection for robust G×E interaction studies. The spike-and-slab priors have been imposed on both the individual and group level to ensure the shrinkage of posterior estimates corresponding to unimportant main and interaction effects to zero exactly. Such a prior leads to the direct sparsity and is superior to the Laplacian types of shrinkage in terms of identification and prediction results (George and McCulloch, 1993; Cassese et al., 2014; Tang et al., 2017; Rocková and George, 2018).

In this study, our objective is to tackle the challenging task of developing a fully Bayesian robust variable selection method for G×E interactions, which has been well motivated by the success of regularization methods (especially those robust ones) in G×E studies and a lack of robust interaction analysis within the Bayesian framework. The significance of the proposed study lies in the following aspects. First, it advances from existing Bayesian G×E studies by incorporating robustness to accommodate data contamination and heavy-tailed distributions in the disease phenotype. Second, on a broader scope, although robust Bayesian quantile regression-based variable selection has been proposed under LASSO, group LASSO, and elastic net, the more complicated sparse group (or bi-level) structure, which is of particular importance in high dimensional data analysis in general (Breheny and Huang, 2009), has not been fully understood yet. We are among the first to develop a robust Bayesian bi-level selection method. Third, unlike existing Bayesian regularized quantile regression methods which build upon the priors under the Laplacian type of shrinkage, we conduct efficient Bayesian regularization on both the individual and group levels by borrowing strength from the spike-and-slab priors. This approach leads to better identification and prediction performance over the competing alternatives, as demonstrated in extensive simulation studies and case studies of NHS data with SNP measurements and TCGA melanoma data with gene expression measurements. To facilitate reproducible research and fast computation using our Markov chain Monte Carlo (MCMC) algorithms, we implement the proposed and alternative methods in C++, which are available from an open-source R package roben (Ren et al., 2020) on CRAN.

2 | DATA AND MODEL SETTINGS

Use subscript i to denote the i th subject. Let $(Y_i, \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{W}_i), (i = 1, \ldots, n)$ be independent and identically distributed random vectors. Y_i is a continuous response variable representing the phenotypic trait. $\boldsymbol{X}_i$ is the p-dimensional vector of G factors. The environmental factors and

clinical covariates are denoted as the k- and q-dimensional vectors $\boldsymbol{E}_i$ and $\boldsymbol{W}_i$, respectively. Considering the following model:

$$\begin{aligned} Y_i &= \sum_{t=1}^{q} \alpha_t W_{it} + \sum_{m=1}^{k} \theta_m E_{im} + \sum_{j=1}^{p} \gamma_j X_{ij} \\ &\quad + \sum_{j=1}^{p} \sum_{m=1}^{k} \zeta_{jm} E_{im} X_{ij} + \epsilon_i \\ &= \sum_{t=1}^{q} \alpha_t W_{it} + \sum_{m=1}^{k} \theta_m E_{im} \\ &\quad + \sum_{j=1}^{p} \left(\gamma_j X_{ij} + \sum_{m=1}^{k} \zeta_{jm} E_{im} X_{ij} \right) + \epsilon_i \\ &= \sum_{t=1}^{q} \alpha_t W_{it} + \sum_{m=1}^{k} \theta_m E_{im} + \sum_{j=1}^{p} \left(U_{ij}^{\top} \beta_j \right) + \epsilon_i, \end{aligned} \tag{1}$$

where α_ts, θ_ms, γ_js, and ζ_{jm}s are the regression coefficients for the clinical covariates, environmental factors, genetic factors, and G×E interactions, correspondingly. We define $\boldsymbol{\beta}_j = (\gamma_j, \zeta_{j1}, \ldots, \zeta_{jk})^{\top} \equiv (\beta_{j1}, \ldots, \beta_{jL})^{\top}$ and $\boldsymbol{U}_{ij} = (X_{ij}, X_{ij}E_{i1} \ldots, X_{ij}E_{ik})^{\top} \equiv (U_{ij1}, \ldots, U_{ijL})^{\top}$, where $L = k + 1$. The coefficient vector $\boldsymbol{\beta}_j$ represents all the main and interaction effects with respect to the jth genetic measurement. The ϵ_i s are random errors. Denote $\boldsymbol{U}_i = \left(\boldsymbol{U}_{i1}^{\top}, \ldots, \boldsymbol{U}_{ip}^{\top}\right)^{\top}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^{\top}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^{\top}$, and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\top}, \ldots, \boldsymbol{\beta}_p^{\top}\right)^{\top}$. The vector $\boldsymbol{\beta}$ is of length $p \times L$. Then model (1) can be written in a more concise form as

$$Y_i = \boldsymbol{W}_i^{\top} \boldsymbol{\alpha} + \boldsymbol{E}_i^{\top} \boldsymbol{\theta} + \boldsymbol{U}_i^{\top} \boldsymbol{\beta} + \epsilon_i . \tag{2}$$

2.1 | Bayesian LAD regression

The LAD regression is well known for its robustness to long-tailed distributions in response. For a Bayesian formulation of LAD regression, we assume that ϵ_is are independently and identically distributed random variables from the Laplace distribution with density

$$f(\epsilon_i \mid \nu) = \frac{\nu}{2} \exp\{-\nu|\epsilon_i|\}, \; i = 1, \ldots, n, \tag{3}$$

where υ^{-1} is the scale parameter of the Laplace distribution. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\top}$. With clinical covariates $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)^{\top}$, environment factors $\boldsymbol{E} = (\boldsymbol{E}_1, \ldots, \boldsymbol{E}_n)^{\top}$, and genetic main effects and G×E interactions $\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)^{\top}$, the likelihood function can be expressed as

$$f(\boldsymbol{Y}\,|\,\boldsymbol{W},\boldsymbol{E},\boldsymbol{U},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\beta},\upsilon)=\prod_{i=1}^{n}\frac{\upsilon}{2}\exp\{-\upsilon|Y_i-\mu_i|\},\tag{4}$$

where $\mu_i=\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\alpha}+\boldsymbol{E}_i^{\mathrm{T}}\boldsymbol{\theta}+\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{\beta}$.

Based on Kozumi and Kobayashi (2011), the Laplace distribution is equivalent to the mixture of an exponential and a scaled normal distribution. Specifically, let z and $\tilde{u}$ be the standard normal and exponential random variables, respectively. If a random variable ϵ follows the Laplace distribution with parameter ν, then it can be represented as follows:

$$\epsilon=\upsilon^{-1}\kappa\sqrt{\tilde{u}}z,\tag{5}$$

where $\kappa=\sqrt{8}$ is a constant. Therefore, the response Y_i can be rewritten as $Y_i=\mu_i+\upsilon^{-1}\kappa\sqrt{\tilde{u}_i}z_i$, where $z_i\sim N(0,\ 1)$ and $\tilde{u}_i\sim\mathrm{Exp}(1)$. Let $u=\upsilon^{-1}\tilde{u}$. Then u follows the exponential distribution $\mathrm{Exp}(\nu)$, where υ is the rate parameter. We thus have the following hierarchical representation of the Laplace likelihood:

$$Y_i=\mu_i+\upsilon^{-\frac{1}{2}}\kappa\sqrt{u_i}z_i,$$

$$u_i\mid\upsilon\overset{ind}{\sim}\mathrm{Exp}(\nu),$$

$$z_i\overset{ind}{\sim}\mathrm{N}(0,\ 1).$$

The conditional distribution of Y_i is normal with mean μ_i and variance $\upsilon^{-1}\kappa^2u_i$. Therefore, the hierarchical representation allows us to express the corresponding likelihood function based on a multivariate normal distribution, making it much easier to formulate a Gibbs sampler that utilizes all the samples for posterior inference. Gibbs sampling is a special case of the Metropolis–Hastings (M-H) algorithm with an acceptance rate for samples being 1. Therefore, it is more efficient in sampling and less computationally intensive than the regular M-H algorithm.

Remark: The Laplace distribution in Bayesian LAD regression can be treated as a special case of the asymmetric Laplace distribution (ALD) in Bayesian quantile regression (Yu and Moyeed, 2001; Yu and Zhang, 2005). The connection between these two distributions is described in Section 1.1 in Web Appendix A.

2.2 | Bayesian sparse group variable selection for G×E interactions

The proposed fully Bayesian sparse group variable selection is motivated by the following considerations. In model (1), the coefficient vector $\boldsymbol{\beta}_j$ corresponds to the main and interaction effects with respect to the jth genetic variant. Whether the genetic variant is associated with the phenotype or not can be determined by whether $\boldsymbol{\beta}_j = \mathbf{0}$. A zero coefficient vector suggests that the variant does not have any effect on the disease outcome. If $\boldsymbol{\beta}_j \neq \mathbf{0}$, then a further investigation on the presence of the main effect, the interaction, or both are of interest, which can be facilitated by examining the nonzero component in $\boldsymbol{\beta}_j$. Therefore, a tailored robust Bayesian variable selection method for G×E studies should accommodate the selection on both group (the entire vector of $\boldsymbol{\beta}_j$) and individual (each component of $\boldsymbol{\beta}_j$) levels at the same time.

In order to impose the robust bi-level sparsity to identify important main and interaction effects, we conduct the decomposition of $\boldsymbol{\beta}_j$ by following the reparameterization from Xu and Ghosh (2015). Here, $\boldsymbol{\beta}_j$ is defined as

$$\boldsymbol{\beta}_j = \boldsymbol{V}_j^{\frac{1}{2}} \boldsymbol{b}_j,$$

where $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{jL})^\top$ and $\boldsymbol{V}_j^{\frac{1}{2}} = \mathrm{diag}\{\omega_{j1}, \ldots, \omega_{jL}\}$, $\omega_{jl} \geq 0$ $(l = 1, \ldots, L)$. To determine whether the j th genetic factor has any effect at all, we conduct group-level selection on $\boldsymbol{b}_j$ by adopting the following multivariate spike-and-slab priors

$$\begin{aligned} &\boldsymbol{b}_j \mid \phi_j^b \overset{ind}{\sim} \phi_j^b \, \mathrm{N}_L(\mathbf{0}, \mathbf{I}_L) + \left(1 - \phi_j^b\right) \delta_0(\boldsymbol{b}_j), \\ &\phi_j^b \mid \pi_0 \overset{ind}{\sim} \mathrm{Bernoulli}(\pi_0), \end{aligned} \tag{6}$$

where $\mathbf{I}_L$ is an identity matrix, $\delta_0(\boldsymbol{b}_j)$ denotes a point mass at $\mathbf{0}_{L \times 1}$ and $\pi_0 \in [0,\ 1]$. We introduce a latent binary indicator variable ϕ_j^b for each group $j (j = 1, \ldots, p)$ to tackle the group-level selection. In particular, when $\phi_j^b = 0$, the coefficient vector b_j has a point mass density at zero and all predictors representing the main and interaction effects in the jth group are excluded from the model, indicating that the jth genetic factor is not associated with the phenotype. On the other hand, when $\phi_j^b = 1$, the components in coefficient vector b_j have nonzero values.

To further determine whether there is an important main genetic effect, G×E interaction, or both, we impose sparsity within the group j by assigning the following spike-and-slab priors on each ω_{jl} ($j = 1, \ldots, p$ and $l = 1, \ldots, L$)

$$\begin{aligned} &\omega_{jl} \mid \phi_{jl}^w \overset{ind}{\sim} \phi_{jl}^w \, \mathrm{N}^+\left(0, s^2\right) + (1 - \phi_{jl}^w) \delta_0(\omega_{jl}), \\ &\phi_{jl}^w \mid \pi_1 \overset{ind}{\sim} \mathrm{Bernoulli}(\pi_1), \end{aligned} \tag{7}$$

where $N^{+}(0, s^2)$ denotes a normal distribution, $N(0, s^2)$, truncated below at 0. When the binary indicator variable $\phi_{jl}^{w} = 0$, ω_{jl} is set to zero by the point mass function $\delta_0(\omega_{jl})$. Within the jth group, when the component $\omega_{jl} = 0$, we have $\beta_{jl} = 0$ and the corresponding U_{jl} is excluded from the model, even when $\boldsymbol{b}_j \neq \mathbf{0}$. This implies that the jth genetic variant does not have the main effect (if $l = 1$) or the interaction effect with the $(l - 1)$ th environment factor (if $l > 1$). The β_{jl} is nonzero if and only if the vector $\boldsymbol{b}_j \neq \mathbf{0}$ and the individual element $\omega_{jl} \neq 0$.

In (6) and (7), π_0 and π_1 control the sparsity on the group and individual levels, respectively. Their values should be carefully tuned. Fixing their values at 0.5 makes the prior essentially weakly informative since equal prior probabilities are given to all the submodels. Instead of fixing π_0 and π_1, we assign conjugate beta priors $\pi_0 \sim \text{Beta}(a_0, b_0)$ and $\pi_1 \sim \text{Beta}(a_1, b_1)$, which can automatically account for the uncertainty in choosing π_0 and π_1. We fix parameters $a_0 = b_0 = a_1 = b_1 = 1$. For computational convenience, we assign a conjugate inverse-Gamma hyperprior on s^2

$$s^2 \sim \text{Inv-Gamma}(1, \eta).$$

η is estimated with the Monte Carlo expectation maximization (EM) algorithm (Park and Casella, 2008; Xu and Ghosh, 2015). For the tth EM update,

$$\eta^{(t)} = \frac{1}{E_{\eta^{(t-1)}}\left[\frac{1}{s^2} \mid \boldsymbol{Y}\right]},$$

where the posterior expectation of $\frac{1}{s^2}$ is estimated from the MCMC samples based on $\eta^{(t-1)}$. To maintain conjugacy, we place a Gamma prior on υ,

$$\upsilon \sim \text{Gamma}(c, d),$$

where c and d are set to small values. The details of full conditional distributions and the Gibbs sampler are available from Web Appendix A.

We term the proposed robust Bayesian sparse group variable selection with spike-and-slab priors as RBSG–SS, with direct competitors robust Bayesian group selection with spike-and-slab priors (denoted as RBG–SS) and robust Bayesian Lasso with spike-and-slab priors (denoted as RBL–SS), and ones without the spike-and-slab priors: RBSG, RBG, and RBL. The nonrobust counterparts are BSG-SS, BG-SS, BL-SS, BSG, BG, and BL, respectively. The summary of abbreviations and definitions of all methods can be found in Web Table 1. All the 12 methods under comparison have been implemented in the C++-based R package roben (Ren et al., 2020) available from CRAN. It is worth mentioning that besides RBSG-SS, the three alternatives RBG-SS, RBL-SS, and RBSG have also been proposed for the first time. The summary of all the methods, including detailed descriptions and the comparison table, is provided in Web Appendix A in the Supporting Information.

Remark: The proposed RBSG–SS respects the weak hierarchy as environmental main effects always retain in the model. Following Wu et al. (2018) and studies alike, we can readily extend the proposed method to accommodate a strong hierarchy (i.e., the genetic main effect also stays in the model once its corresponding G×E interaction is identified) under the spike-and-slab priors by penalizing interaction effects on both the group and individual level while only regularizing the main genetic effect on the group level. We have not pursued imposing the strong hierarchy as it does not lead to improved prediction and identification performance over the RBSG–SS.

3 | SIMULATION

We comprehensively evaluate the proposed and alternative methods through simulation studies. Under all the settings, the responses are generated from model (1) with $n = 500$, $q = 3$, $p = 100$, and $k = 5$, which leads to a total dimension of 608 with 105 main effects, 500 interactions, and three additional clinical covariates. The genetic main effects and G×E interactions form 100 groups with group size $L = 6$. We consider five error distributions for $\epsilon_i s$: N(0, 1)(Error 1), Laplace(μ, b) with the mean $\mu = 0$, and the scale parameter $b = 2$ (Error 2), $10\%\text{Laplace}(0,1) + 90\%\text{Laplace}(0, \sqrt{5})$ (Error 3), t-distribution with 2 degrees of freedom ($t(2)$) (Error 4), LogNormal(0,1) (Error 5). All of them are heavy-tailed distributions except the first one.

We assess the performance in terms of identification and prediction accuracy. For methods incorporating spike-and-slab priors, we adopt the median probability model (MPM) (Barbieri and Berger, 2004; Xu and Ghosh, 2015) to identify important effects. In particular, for the proposed RBSG–SS, we define $\phi_{jl} = \phi_j^b \phi_{jl}^w$ for the lth predictor in the jth group. At the gth MCMC iteration, this predictor is included in the model if the indicator $\phi_{jl}^{(g)}$ is 1. Suppose we have collected G posterior samples from the MCMC after burn-ins, then the posterior probability of including the lth predictor from the jth group in the final model is

$$p_{jl} = \hat{\pi}(\phi_{jl} = 1 \mid y) = \frac{1}{G} \sum_{g=1}^{G} \phi_{jl}^{(g)},$$
$$j = 1, \ldots, p \text{ and } l = 1, \ldots, L. \tag{8}$$

A higher posterior inclusion probability p_{jl} can be interpreted as a stronger piece of empirical evidence that the corresponding predictor has a nonzero coefficient and is associated with the phenotype. The MPM model is defined as the model consisting of predictors with at least $\frac{1}{2}$ posterior inclusion probability. When the goal is to select a single model, Barbieri and Berger (2004) recommend using MPM because of its optimal prediction performance. Meanwhile, the 95% credible interval (95% CI) (Li et al., 2015) is adopted for methods without spike-and-slab priors.

Prediction performance is evaluated using the mean prediction errors on an independently generated testing dataset under the same data-generating model over 100 replicates. For all

robust approaches, the prediction error is defined as mean absolute deviations (MAD). MAD can be computed as $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. The prediction error for nonrobust ones is defined as the mean squared error (MSE), that is, $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

The G factors are simulated in the following four examples(settings). In the first example, a gene expression matrix with $n = 500$ and $p = 100$ has been generated from a multivariate normal distribution with marginal mean 0, marginal variance 1, and an autoregression correlation structure ($\rho = 0.3$). In the second example, the SNP data are obtained by dichotomizing the gene expression values (from the first setting) at the first and third quartiles, with the three-level (0,1,2) for genotypes (aa, Aa, AA), respectively. In the third setting, the SNP data are simulated under a pairwise linkage disequilibrium (LD) structure. Let the minor allele frequencies (MAFs) of two neighboring SNPs with risk alleles A and B be r_1 and r_2, respectively. The frequencies of four haplotypes are as $p_{AB} = r_1r_2 + \delta$, $p_{ab} = (1 - r_1)(1 - r_2) + \delta$, $p_{Ab} = r_1(1 - r_2) - \delta$, and $p_{aB} = (1 - r_1)r_2 - \delta$, where δ denotes the LD. Assuming Hardy–Weinberg equilibrium and given the allele frequency for A at locus 1, we can generate the SNP genotype (AA, Aa, aa) from a multinomial distribution with frequencies $(r_1^2, 2r_1(1 - r_1), (1 - r_1)^2)$. The genotypes at locus 2 can be simulated according to the conditional genotype probability matrix in Cui et al. (2008). We have $\delta = r_p\sqrt{r_1(1 - r_1)r_2(1 - r_2)}$ with MAFs 0.3 and pairwise correlation $r_p = 0.6$. In the last example, we consider a more practical correlation structure by extracting the first 100 SNPs from the NHS data analyzed in the case study, so the correlation is based on the real data. For each simulation replicate, we randomly sample 500 subjects from the dataset.

For E factors, five continuous variables are generated from a multivariate normal distribution with marginal mean 0, marginal variance 1, and auto regression (AR) correlation structure with $\rho = 0.5$. We then dichotomize one of them at 0 to create a binary E factor. Besides, we simulate three clinical covariates from a multivariate normal distribution and AR correlation structure with $\rho = 0.5$ and dichotomize one of them at 0 to create a binary variable.

For the clinical covariates and environmental main effects, the coefficients α_ts and θ_ms are generated from Uniform [0.8, 1.5]. For genetic main effect and G×E interactions, we randomly selected 25 β_{jl}s in nine groups to have nonzero values that are generated from Uniform [0.3, 0.9]. All other β_{jl}s are set to zeros.

We have collected the posterior samples from the Gibbs sampler running 15,000 iterations while discarding the first 7500 samples as the burn-in. The Bayesian estimates are calculated using the posterior medians. Simulation results for the gene expression data in Example 1 are tabulated in Tables 1 and 2. We can observe that the performance of methods that adopt spike-and-slab priors in Table 1 is consistently better than methods without spike-and-slab priors in Table 2. Although methods without spike-and-slab priors have slightly lower FPs than their counterparts with spike-and-slab priors under some error distributions, they tend to have much lower TPs and higher prediction errors under all the error distributions. For example, under Error 2, RBSG identifies 14.48 (SD 2.04) out of the 25 true positives, much lower than the true positives of 21.66 (SD 1.72) from RBSG–SS. Meanwhile, its

false positives 0.64 (SD 0.85) is only slightly lower than the FP of RBSG–SS (1.32 (SD 1.33)). The prediction error of RBSG, 2.57 with an SD of 0.11, is also inferior to that of the RBSG–SS (2.15 (SD 0.10)). Such an advantage can also be observed by comparing other methods in Table 1 with their counterparts (without spike-and-slab priors) from Table 2.

Among all the methods with spike-and-slab priors, as shown in Table 1, the proposed RBSG–SS has the best performance in both identification and prediction in the presence of data contamination and heavy-tailed errors. Under the mixture Laplace error (Error 3), RBSG–SS identifies 21.28 (SD 2.24) true positives, with a small number of false positives, 1.48 (SD 1.34). RBG–SS has a true positive of 24.80 (SD 0.73); however, the number of false positives, 30.64 (SD 4.23), is much higher. This is due to the fact that RBG–SS only conducts a group-level selection and does not impose the within-group sparsity. Compared to RBSG–SS, RBL–SS ignores the group structure, leading to fewer true positives of 18.14 (SD2.68). In terms of prediction, RBSG–SS has the smallest L1 error, 2.29 (0.12), among all the three robust methods with spike-and-slab priors. Although the difference in prediction error between RBSG–SS and RBG–SS is not distinct, considering the much smaller number of false-positive main and interaction effects, we can fully observe the advantage of RBSG–SS over RBG–SS in prediction.

Moreover, a cross-comparison between the robust and nonrobust methods further demonstrates the necessity of developing robust Bayesian methods. For instance, under the error of t distribution with 2 degrees of freedom (Error 4), RBSG–SS has identified 23.80 (SD 1.30) true main and interaction effects with only 0.53 (SD 0.86) false positives. Its direct nonrobust competitor, BSG–SS, leads to a true positive of 16.20 (SD 6.45) with 3.73 (SD 4.61) false effects. The superior performance of RBSG–SS over the other two nonrobust methods, BG–SS and BL–SS, is also clear. Although a comparison between the prediction errors of robust and nonrobust methods is not straightforward as the two are computed under the L1 and least-square errors, the identification results convincingly suggest the advantage of robust methods over nonrobust ones. The graphical representation for Tables 1 and 2 using TP, FP, as well as other metrics such as the Matthew correlation coefficient and F-score, can be found in Web Appendix D.

Similar patterns have been observed in Web Table 3–8 for Examples 2, 3, and 4, respectively, in Web Appendix B. Overall, based on the investigations of all the methods through comprehensive simulation studies, we can establish the advantage of conducting robust Bayesian bi-level selection incorporating spike-and-slab priors.

We demonstrate the sensitivity of RBSG–SS to the choice of the hyperparameters for π_0, π_1, and ν in Web Appendix B. The results are tabulated in Web Table 9, showing that the MPM model is insensitive to different specifications of the hyperparameters. The convergence of the MCMC chains and the computational cost are also assessed. The results provided in Web Figure 2 and Web Table 10 show that the convergence of the proposed Gibbs sampler can be achieved with a reasonable computation time.

We have performed additional simulation studies to further investigate and demonstrate the advantage of RBSG-SS over the alternatives with spike-and-slab priors. The rest of

the methods are not included in the new simulations due to their inferior performance. Specifically, we have reported the identification performance for main and interaction effects separately, and the estimation accuracy for zero-effects and nonzero effects separately for Examples 1–4. Example 5, the additional data-generating setting using the TCGA SKCM data analyzed in the case study, has also been considered. The results are tabulated in Web Appendix C. In addition, instead of merely using positive nonzero coefficients, we have considered randomly assigning half of the nonzero effects from Uniform[0.3,0.9] and the other half from Uniform[−0.9, −0.3]. The results are provided in Web Tables 63–67 in Web Appendix D, demonstrating the advantage of RBSG–SS. Besides, we have conducted additional simulations to compare the performance of the proposed method with three frequentist methods, Quantile Lasso (QL), quantile group Lasso (QGL), and sparse group Lasso (SGL). The results given in Web Appendix D show that RBSG-SS outperforms the frequentist methods in both identification and prediction. Overall, RBSG-SS has shown superior performance to the alternatives in all studies.

Remark:

Identification of important effects by robust methods without spike-and-slab priors has been conducted using the 95% credible interval, which has been widely adopted in existing regularized Bayesian quantile regression studies based on the asymmetric Laplace (AL) likelihood. As the AL likelihood is usually not the true data-generating likelihood, Yang et al. (2016) has proposed an adjustment on posterior variance to construct asymptotically valid credible intervals, which improves the performance. We have compared the proposed RBSG-SS and other spike-and-slab based methods with Bayesian quantile LASSO implemented by using Package bayesQR. The results in Web Table 49 have shown that the method without the spike-and-slab prior consistently has inferior performance. Since by construction, Bayesian shrinkage using the spike-and-slab priors leads to the sparsity directly and outperforms the counterparts that depend on credible intervals to impose the sparsity, we have not implemented such a correction on the alternative methods without spike-and-slab priors.

4 | REAL DATA ANALYSIS

4.1 | Nurses' Health Study data

NHS is one of the largest investigations into the risk factors for major chronic diseases in women. As part of the Gene Environment Association Studies initiative (GENEVA), the NHS provides SNP genotypes data as well as detailed information on dietary and lifestyle variables. Obesity level is one of the most important risk factors for type 2 diabetes mellitus (T2D), a chronic disease due to both genetic and environmental factors. In this study, we analyze the NHS type 2 diabetes data to identify main and interaction effects associated with obesity. We use weight as the response and focus on SNPs on chromosome 10. We consider five environmental factors, including the total physical activity (act), glycemic load (gl), cereal fiber intake (ceraf), alcohol intake (alcohol), and a binary indicator of whether an individual has a history of high cholesterol (chol). All these environmental exposures have been suggested to be associated with obesity and diabetes (Hu et al., 2001). In addition, we include three clinical covariates: height, age, and a binary indicator of whether an individual

has a history of hypertension (hbp). In the NHS study, about half of the subjects are diagnosed with type 2 diabetes and the other half are controls without the disease. We only use healthy subjects in this study. After cleaning the data through matching phenotypes and genotypes, removing SNPs with MAF less than 0.05 or deviation from the Hardy–Weinberg equilibrium, the working dataset contains 1732 subjects with 35,099 SNPs.

For computational convenience, prescreening can be conducted to reduce the feature space to a more attainable size for variable selection. For example, Li et al. (2015) and Wu et al. (2014) use the single SNP analysis to filter SNPs in a Genome wide association study before downstream analysis. In this study, we use a marginal linear model with weight as the response variable to evaluate the penetrance effect of a variant under environmental exposure. The marginal linear model uses a group of genetic main effect and G×E interactions corresponding to an SNP as the predictors and tests whether this SNP has any effect, main, or G×E interaction. The SNPs with p-values less than a certain cutoff (0.001) for any main or interaction effect from the test are kept. Two-hundred and fifty-three SNPs pass the screening.

The proposed approach RBSG-SS identifies 22 main SNP effects and 45 G×E interactions. The detailed estimation results are provided in Web Table 11. Genes corresponding to these SNPs are identified, and discussions on their biological functionalities are provided in Web Appendix B. Literature mining suggests that these genes and interactions may have important implications in obesity, which may provide support to the validity of RBSG-SS. The convergence of the MCMC chains is assessed, and the results can be found in Web Appendix D.

In addition to the proposed approach, we also conduct analysis using the alternatives RBL–SS, BSG–SS, and BL–SS. As other alternative methods show inferior performance in simulation, they are not considered in real data analysis. Detailed estimation results are provided in Web Tables 12–14 in Web Appendix B. In Table 3, we provide the numbers of identified main and interaction effects with pairwise overlaps to show the difference in terms of identification between the proposed method and the others. To further investigate the biological similarity of the identified genes, we also conduct the gene ontology (GO) analysis. We can find an obvious difference between the proposed RBSG–SS and the three alternatives. The GO analysis results are provided in Web Figure 3 from Web Appendix B.

With real data, it is difficult to assess the selection accuracy objectively. The prediction performance may provide additional information to the selection results. Following Yan and Huang (2012) and Li et al. (2015), we refit the models identified by RBSG–SS and RBL–SS using the robust Bayesian Lasso and refit the models selected by BSG–SS and BL–SS using the Bayesian Lasso. For robust methods, the prediction mean absolute deviations (PMAD) are computed based on the posterior median estimates, while the prediction mean squared errors, or PMSEs, are computed for nonrobust methods. Other procedures, such as posterior predictive checking, are also potentially applicable but it requires further investigation before their application in the current data setting. To facilitate the cross-comparison in prediction between robust and nonrobust methods, we also computed PMAD for nonrobust methods and PMSE for robust methods. We evaluated the prediction performance based

on multiple data splitting. For each split, one-fifth of the subjects are randomly selected as the testing set and the rest of the data is used as the training set. Both PMAD and PMSE are computed for each method on the testing set. The mean and standard deviation of prediction errors over 100 splittings are recorded. For robust methods RBSG-SS and RBL-SS, the PMADs are 8.61 (SD 0.30) and 8.84 (SD 0.34) and the PMSEs are 122.41 (SD 10.38) and 128.36 (SD 11.17), respectively. For nonrobust methods BSG-SS and BL-SS, the PMADs are 8.95 (SD 0.34) and 9.22 (SD 0.36) and the PMSEs are 131.08 (SD 11.17) and 139.03 (SD 11.46), respectively. RBSG-SS has the smallest PMADs and PMSEs than the alternatives.

4.2 | TCGA skin cutaneous melanoma data

We analyze the TCGA SKCM data. TCGA is a collaborative effort supported by the National Cancer Institute and the National Human Genome Research Institute and has published high-quality clinical, environmental, as well as multiomics data. For this study, we use the level-3 gene expression data of SKCM downloaded from the cBio Cancer Genomics Portal (Cerami et al., 2012). Our goal is to identify genes that have a genetic main effect or G×E interaction effects on the Breslow's thickness, an important prognostic variable for SKCM (Marghoob et al., 2000). The log-transformed Breslow's depth is used as the response variable and four E factors, age, American Joint Committee on Cancer pathologic tumor stage, gender, and Clark level are considered. Data are available on 294 subjects and 20,531 gene expressions. We adopt the same screening method used in the first case study to select 109 genes for further analysis.

The proposed RBSG–SS identifies 16 main effects and 32 G×E interactions. The detailed estimation results are available from Web Table 15. For the identified genes, published literature provides independent evidence of their associations with cutaneous melanoma. Discussions are given in Web Appendix B. Details on diagnostic analysis of the convergence of the MCMC chains can be found in Web Appendix D.

Analysis is also conducted via the three alternative methods, and the results are summarized in Table 3. Detailed estimation results are provided in Web Tables 16–18. Again, the proposed RBSG–SS identifies different sets of main and interaction effects from the rest. We further investigate the biological similarity of the identified genes by GO analysis (Web Figure 3), which suggests an obvious difference. Prediction performance is also evaluated.

In the multiple data splitting, the robust methods RBSG-SS and RBL-SS have PMADs as 0.66 (SD 0.08) and 0.82 (SD 0.11), and PMSEs as 0.87 (SD 0.34) and 1.15 (0.43), respectively. The nonrobust methods BSG-SS and BL-SS have PMADs as 0.79 (SD 0.09) and 0.87 (SD 0.12) and PMSEs as 0.96 (SD 0.29) and 1.25 (SD 0.46), respectively. RBSG-SS outperforms the alternatives in terms of both PMADs and PMSEs.

Remark: Methods incorporating spike-and-slab priors lead to the posterior inclusion probabilities that can be used as a measure for quantifying uncertainty for variable selection. The inclusion probabilities of both case studies are provided in Section 3.3 from Web Appendix C. A higher posterior inclusion probability indicates a more stable selection result

for the corresponding effects and suggests the stronger empirical evidence that it has a nonzero coefficient.

5 | DISCUSSION

In this study, we have developed robust Bayesian variable selection methods for gene–environment interaction studies. The robustness comes from the Bayesian formulation of LAD regression which corresponds to the symmetric Laplace distribution, a special scenario of the ALD in Bayesian quantile regression. The symmetric assumption has been adopted in our study since the distributions of residuals shown in Figure 1 are not extreme, and RBSG-SS with a quantile level at 50% works well in such a case as partially demonstrated by its ideal performance under the skewed log-normal error in simulation. Nevertheless, determining the optimal (combination of) quantiles is an interesting topic for future research. In G×E studies, the demand for robustness arises in (1) heavy-tailed distribution and data contamination in the response and/or (2) in the predictors, as well as (3) model misspecification. We have focused on the first case, which is frequently encountered in practice. Investigations of the robust Bayesian methods accommodating the other two cases are interesting and will be pursued in the future.

Recently, penalization has emerged as a powerful tool for dissecting G×E interactions (Zhou et al., 2021). Our literature review suggests that Bayesian variable selection methods, although tightly related to penalization, have not been fully explored for interaction analyses, let alone the robust ones. We are among the first to conduct robust G×E analysis within the Bayesian framework. The proposed Bayesian LAD sparse group selection is not only specifically tailored for G×E studies but also generally applicable for problems incorporating the bi-level structure in a broader context, such as simultaneous selection of prognostic genes and pathways (Liu et al., 2019). The spike-and-slab priors have been incorporated to further improve identification and prediction performance. As a by-product, the Bayesian LAD LASSO and group LASSO, both with spike-and-slab priors, have also been investigated for the first time. The computational feasibility and reproducibility of the Gibbs samplers are guaranteed by the R package roben (Ren et al., 2020) available on CRAN, with the core modules of the MCMC algorithms developed in C++.

In G×E studies, the form of interaction effects can be linear, nonlinear, and both linear and nonlinear, resulting in parametric (Wu et al., 2018; Zhou et al., 2019), nonparametric (Li et al., 2015; Wu et al., 2018) and semiparametric variable selection methods (Wu et al., 2014, 2015; Ren et al., 2020) to dissect G×E interactions, respectively. The proposed study can be potentially generalized to these studies within a robust Bayesian framework. For example, variable selection for multiple semiparametric G×E studies can be formulated as a combination of individual- and group-level selection problems, where the robust Bayesian methods based on sparse group, group and individual level selection are directly applicable. The proposed robust Bayesian framework has paved the way for future investigations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank the editor, associate editor, and two anonymous reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. We also thank Dr. Travis S. Johnson for his careful proofreading of our manuscript. This work was partially supported by an Innovative Research Award from the Johnson Cancer Research Center at Kansas State University and the National Institutes of Health (NIH) grant R01 CA204120.

Funding information

National Institutes of Health, Grant/Award Number: R01 CA204120; The Johnson Cancer Research Center at Kansas State University, Grant/Award Number: The Innovative Research Award

DATA AVAILABILITY STATEMENT

The Nurses' Health data are not publicly available. Researchers may apply from the Database of Genotype and Phenotype (dbGaP) at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1 through accession number phs000091.v2.p1. The Cancer Genome Atlas Skin Cutaneous Melanoma data are openly available via https://portal.gdc.cancer.gov/.

REFERENCES

Barbieri MM and Berger JO (2004) Optimal predictive model selection. The Annals of Statistics, 32, 870–897.

Breheny P and Huang J (2009) Penalized methods for bi-level variable selection. Statistics and Its Interface, 2, 369–380. [PubMed: 20640242]

Cassese A, Guindani M, Tadesse MG, Falciani F and Vannucci M (2014) A hierarchical Bayesian model for inference of copy number variants and their association to gene expression. The Annals of Applied Statistics, 8, 148–175. [PubMed: 24834139]

Centers for Disease Control and Prevention (2020) National Diabetes Statistics Report.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. (2012) The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discovery, 2, 401–404. [PubMed: 22588877]

Cui Y, Kang G, Sun K, Qian M, Romero R and Fu W (2008) Gene-centric genomewide association study via entropy. Genetics, 179, 637–650. [PubMed: 18458106]

Fan J and Lv J (2010) A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20, 101–148. [PubMed: 21572976]

George EI and McCulloch RE (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88, 881–889.

Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, et al. (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. New England Journal of Medicine, 345, 790–797. [PubMed: 11556298]

Hunter DJ (2005) Gene–environment interactions in human diseases. Nature Reviews Genetics, 6, 287–298.

Kozumi H and Kobayashi G (2011) Gibbs sampling methods for Bayesian quantile regression. Journal of Statistical Computation and Simulation, 81, 1565–1578.

Li J, Wang Z, Li R and Wu R (2015) Bayesian group Lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. The Annals of Applied Statistics, 9, 640–664. [PubMed: 26478762]

Li Q, Xi R and Lin N (2010) Bayesian regularized quantile regression. Bayesian Analysis, 5, 533–556.

Liu W, Ouyang S, Zhou Z, Wang M, Wang T, Qi Y, et al. (2019) Identification of genes associated with cancer progression and prognosis in lung adenocarcinoma: Analyses based on microarray from

oncomine and the Cancer Genome Atlas databases. Molecular Genetics & Genomic Medicine, 7, e00528. [PubMed: 30556321]

Marghoob AA, Koenig K, Bittencourt FV, Kopf AW and Bart RS (2000) Breslow thickness and Clark level in melanoma. Cancer, 88, 589–595. [PubMed: 10649252]

Mukherjee B, Ahn J, Gruber SB and Chatterjee N (2011) Testing gene–environment interaction in large-scale case-control association studies: possible choices and comparisons. American Journal of Epidemiology, 175, 177–190. [PubMed: 22199027]

Park T and Casella G (2008) The Bayesian Lasso. Journal of the American Statistical Association, 103, 681–686.

Ren J, Du Y, Li S, Ma S, Jiang Y and Wu C (2019) Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. Genetic Epidemiology, 43, 276–291. [PubMed: 30746793]

Ren J, Zhou F, Li X, Chen Q, Zhang H, Ma S, et al. (2020) Semiparametric Bayesian variable selection for gene-environment interactions. Statistics in Medicine, 39, 617–638. [PubMed: 31863500]

Ren J, Zhou F, Li X and Wu C (2020) roben: Robust Bayesian Variable selection for gene-environment interactions. R package version 0.1.0.

Rocková V and George EI (2018) The spike-and-slab lasso. Journal of the American Statistical Association, 113, 431–444.

Simonds NI, Ghazarian AA, Pimentel CB, Schully SD, Ellison GL, Gillanders EM, et al. (2016) Review of the gene-environment interaction literature in cancer: What do we know? Genetic Epidemiology, 40, 356–365. [PubMed: 27061572]

Tang Z, Shen Y, Zhang X and Yi N (2017) The spike-and-slab Lasso generalized linear models for prediction and associated genes detection. Genetics, 205, 77–88. [PubMed: 27799277]

Wu C, Cui Y and Ma S (2014) Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. Statistics in Medicine, 33, 4988–4998. [PubMed: 25146388]

Wu C, Jiang Y, Ren J, Cui Y and Ma S (2018) Dissecting gene–environment interactions: A penalized robust approach accounting for hierarchical structures. Statistics in Medicine, 37, 437–456. [PubMed: 29034484]

Wu C and Ma S (2015) A selective review of robust variable selection with applications in bioinformatics. Briefings in Bioinformatics, 16, 873–883. [PubMed: 25479793]

Wu C, Shi X, Cui Y and Ma S (2015) A penalized robust semiparametric approach for gene–environment interactions. Statistics in Medicine, 34, 4016–4030. [PubMed: 26239060]

Wu C, Zhang Q, Jiang Y and Ma S (2018) Robust network-based analysis of the associations between (epi)genetic measurements. Journal of Multivariate Analysis, 168, 119–130. [PubMed: 30983643]

Wu C, Zhong P-S and Cui Y (2018) Additive varying-coefficient model for nonlinear gene–environment interactions. Statistical Applications in Genetics and Molecular Biology, 17, 20170008.

Xu X and Ghosh M (2015) Bayesian variable selection and estimation for group Lasso. Bayesian Analysis, 10, 909–936.

Yan J and Huang J (2012) Model selection for Cox models with time-varying coefficients. Biometrics, 68, 419–428. [PubMed: 22506825]

Yang Y, Wang HJ and He X (2016) Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. International Statistical Review, 84, 327–344.

Yu K and Moyeed RA (2001) Bayesian quantile regression. Statistics and Probability Letters, 54, 437–447.

Yu K and Zhang J (2005) A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics—Theory and Methods, 34, 1867–1879.

Zhou F, Ren J, Li G, Jiang Y, Li X, Wang W, et al. (2019) Penalized variable selection for lipid-environment interactions in a longitudinal lipidomics study. Genes, 10, 1002. [PubMed: 31816972]

Zhou F, Ren J, Lu X, Ma S and Wu C (2021) Gene–environment interaction: a variable selection perspective. Epistasis. Methods in Molecular Biology, 2212, 191–223. [PubMed: 33733358]

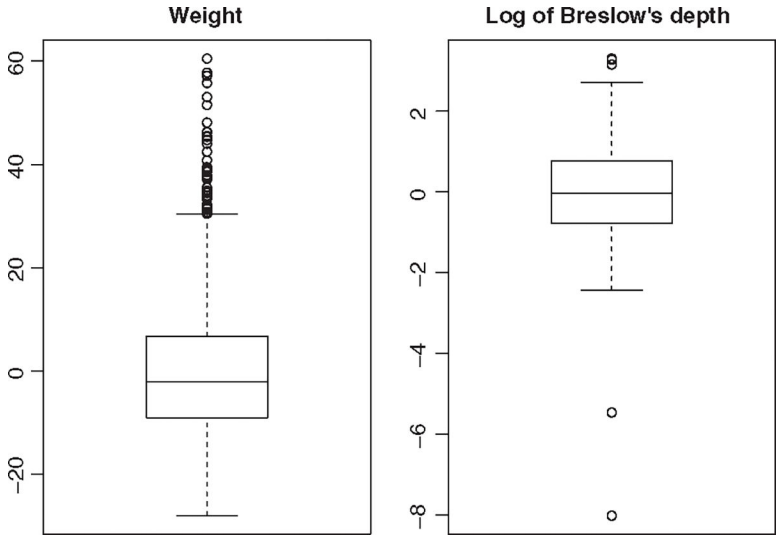

FIGURE 1.

Distribution of the residuals for the NHS (left) and SKCM (right) data

TABLE 1

Simulation results in Example 1 for methods with spike–and–slab priors. (n, q, k, p) = (500, 3, 5, 100). mean (SD) of true positives (TP), false positives (FP), and prediction errors (Pred) based on 100 replicates

		RBSG–SS	RBG–SS	RBL–SS	BSG–SS	BG–SS	BL–SS
Error 1	TP	24.97 (0.18)	25.00 (0.00)	24.93 (0.25)	24.97 (0.18)	25.00 (0.00)	24.93 (0.25)
	FP	1.30 (1.24)	29.60 (2.42)	1.30 (1.44)	0.47 (0.68)	29.00 (0.00)	0.43 (0.73)
	Pred	0.83 (0.03)	0.86 (0.03)	0.84 (0.04)	1.07 (0.07)	1.13 (0.07)	1.08 (0.08)
Error 2	TP	21.66 (1.72)	24.84 (0.55)	18.58 (2.14)	19.98 (1.95)	24.58 (0.86)	15.54 (2.04)
	FP	1.32 (1.33)	30.96 (4.27)	1.62 (1.64)	1.82 (1.53)	30.98 (4.83)	0.92 (0.94)
	Pred	2.15 (0.10)	2.17 (0.09)	2.24 (0.12)	9.32 (0.97)	8.98 (0.79)	10.09 (1.08)
Error 3	TP	21.28 (2.24)	24.80 (0.73)	18.14 (2.68)	19.00 (2.61)	24.40 (1.09)	14.24 (2.39)
	FP	1.48 (1.34)	30.64 (4.23)	1.42 (1.63)	2.04 (1.73)	30.20 (4.73)	1.18 (1.16)
	Pred	2.29 (0.12)	2.32 (0.11)	2.41 (0.12)	11.11 (1.12)	10.59 (0.95)	12.02 (1.12)
Error 4	TP	23.80 (1.30)	24.93 (0.37)	21.80 (1.94)	16.20 (6.45)	21.83 (5.24)	12.53 (5.79)
	FP	0.53 (0.86)	29.47 (2.56)	0.20 (0.41)	3.73 (4.61)	35.77 (23.92)	1.93 (2.49)
	Pred	1.50 (0.14)	1.52 (0.13)	1.53 (0.14)	12.48 (6.56)	12.34 (7.27)	13.35 (6.72)
Error 5	TP	24.33 (0.76)	25.00 (0.00)	22.93 (1.20)	22.93 (1.26)	25.00 (0.00)	18.00 (2.17)
	FP	0.26 (0.45)	29.00 (0.00)	0.13 (0.35)	4.30 (3.40)	34.80 (8.11)	1.23 (1.55)
	Pred	1.16 (0.10)	1.18 (0.10)	1.18 (0.10)	4.75 (1.24)	4.78 (1.23)	5.18 (1.34)

TABLE 2

Simulation results in Example 1 for methods without spike–and–slab priors. (n, q, k, p) = (500, 3, 5, 100). mean(sd) of true positives (TP), false positives (FP), and prediction errors (Pred) based on 100 replicates

		RBSG	RBG	RBL	BSG	BG	BL
Error 1	TP	21.87 (1.38)	24.67 (0.76)	21.97 (1.40)	22.93 (1.34)	24.93 (0.37)	23.07 (1.23)
	FP	2.63 (1.94)	55.33 (15.76)	3.07 (2.35)	2.43 (1.77)	83.47 (20.07)	11.20 (4.34)
	Pred	1.15 (0.05)	1.37 (0.06)	1.15 (0.05)	1.73 (0.12)	2.29 (0.19)	2.21 (0.17)
Error 2	TP	14.48 (2.04)	23.06 (1.96)	14.42 (2.12)	15.18 (2.06)	24.02 (1.48)	15.48 (2.30)
	FP	0.64 (0.85)	32.26 (7.41)	0.74 (0.88)	2.20 (1.55)	85.78 (20.06)	14.06 (4.41)
	Pred	2.57 (0.11)	2.85 (0.13)	2.57 (0.11)	12.43 (1.15)	15.92 (1.68)	16.55 (1.69)
Error 3	TP	13.74 (2.65)	22.52 (2.38)	13.80 (2.66)	14.30 (2.70)	23.92 (1.37)	14.62 (2.69)
	FP	0.68 (0.68)	34.24 (8.93)	0.80 (0.83)	2.74 (1.48)	97.40 (19.78)	15.96 (4.30)
	Pred	2.71 (0.12)	3.00 (0.14)	2.71 (0.12)	14.36 (1.35)	18.52 (1.70)	19.25 (1.84)
Error 4	TP	16.90 (3.12)	21.83 (3.04)	16.90 (3.36)	11.70 (5.86)	20.70 (5.74)	12.07 (5.44)
	FP	0.33 (0.48)	27.97 (8.48)	0.27 (0.45)	3.10 (2.64)	88.50 (28.58)	14.83 (5.52)
	Pred	1.85 (0.15)	2.10 (0.17)	1.85 (0.15)	16.25 (9.88)	22.78 (17.05)	24.20 (18.91)
Error 5	TP	16.26 (2.28)	23.42 (2.01)	16.42 (2.16)	13.80 (3.37)	23.24 (2.25)	14.24 (3.05)
	FP	0.32 (0.62)	29.38 (7.54)	0.32 (0.65)	3.00 (2.14)	94.72 (27.12)	16.26 (4.84)
	Pred	2.20 (0.14)	2.49 (0.17)	2.21 (0.14)	15.94 (4.43)	20.73 (5.12)	21.66 (5.48)

TABLE 3

Identification results for real data analysis. The numbers of main G effects and interactions identified by different approaches and their overlaps

	Main G effects				Interactions			
	RBSG-SS	**RBL-SS**	**BSG-SS**	**BL-SS**	**RBSG-SS**	**RBL-SS**	**BSG-SS**	**BL-SS**
NHS								
RBSG-SS	22	20	16	13	45	21	17	10
RBL-SS		29	20	16		39	14	14
BSG-SS			29	25			34	22
BL-SS				27				42
SKCM								
RBSG-SS	16	10	14	13	32	11	18	10
RBL-SS		17	12	14		33	15	24
BSG-SS			22	15			29	14
BL-SS				20				33