frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fgene.2021.667074

Identifying Gene–Environment Interactions With Robust Marginal Bayesian Variable Selection

Xi Lu[1], *Kun Fan*[1], *Jie Ren*[2] *and Cen Wu*[1]*

[1] *Department of Statistics, Kansas State University, Manhattan, KS, United States,* [2] *Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, United States*

In high-throughput genetics 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, we have implemented the Gibbs sampler based on Markov Chain Monte Carlo (MCMC). The proposed method outperforms a number of alternatives in extensive simulation studies. The utility of the marginal robust Bayesian variable selection method has been further demonstrated in the case studies using data from the Nurse Health Study (NHS). 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, markov chain monte carlo method

OPEN ACCESS

Edited by:
Qi Yan,
Columbia University, United States

Reviewed by:
Rong Zhang,
Amgen, United States
Zilin Li,
Harvard University, United States

****Correspondence:***
Cen Wu
wucen@ksu.edu

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *11 February 2021*
Accepted: *13 July 2021*
Published: *08 December 2021*

Citation:
Lu X, Fan K, Ren J and Wu C (2021)
Identifying Gene–Environment
Interactions With Robust Marginal
Bayesian Variable Selection.
Front. Genet. 12:667074.
doi: 10.3389/fgene.2021.667074

1. INTRODUCTION

The risk and progression of complex diseases including cancer, asthma and type 2 diabetes are associated with the coordinated functioning of genetic factors, the environmental (and clinical) factors, as well as their interactions (Hunter, 2005; Von Mutius, 2009; Cornelis and Hu, 2012; Simonds et al., 2016). The identification of important gene–environment(G×E) interactions leads to novel insight in dissecting the genetic basis of complex diseases in addition to the main effects of genetic and environmental factors. In the last two decades, searching for the important G×E interactions has been extensively conducted based on genetic association studies (Cordell and Clayton, 2005; Wu et al., 2012). One representative example is the genome-wide association study (GWAS), where the statistical significance of interaction between the environmental exposure and the genetic variant has been marginally assessed one at a time across the whole genome. Important findings are evidenced by genome-wide significant *p*-values after adjusting for multiple comparisons.

Recently, substantial efforts have been devoted to novel penalized variable selection methods for G×E studies (Zhou et al., 2021). In particular, marginal penalization has achieved very competitive performances with the aforementioned significance-based G×E analysis (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020). For example, within the framework of maximum rank correlation, Shi et al. (2014) has developed a penalization method robust to outliers and model misspecification in determining important G×E interactions one at a time. Zhang et al. (2020) has imposed hierarchical structure between the main effects and interactions in marginal identification of G×E interactions using regularization. Despite success, these studies have limitations. First, as a common tuning parameter is demanded for all the marginal models, its selection requires pooling all genes together to conduct a joint model-based cross-validation. While such a strategy is not rare, it seems not in favor of the marginal nature of the proposed G×E studies. Second, a rigorous measure to quantify uncertainty is not available. Zhang et al. (2020) has constructed 95% confidence intervals based on the observed occurrence index (OOI) values (Huang and Ma, 2010); nevertheless, this measure has been used to demonstrate stability of identified effects rather than quantifying uncertainty of penalized estimates.

These limitations have motivated us to consider Bayesian analyses. In literature, Bayesian variable selection methods have been developed for G×E analysis in multiple studies (Zhou et al., 2021). For example, with indicator model selection, Liu et al. (2015) has imposed hierarchical Bayesian variable selection for linear G×E interactions. Li et al. (2015) has proposed a Bayesian group LASSO to identify non-linear interactions in nonparametric varying coefficient models. Ren et al. (2020) has further incorporated selection of linear and nonlinear G×E interactions simultaneously while accounting for structured identification in the Bayesian adaptive shrinkage framework. All these fully Bayesian methods can efficiently provide uncertainty quantification based on the posterior samples from MCMC. Nevertheless, our limited literature mining shows that none of the marginal Bayesian variable selection methods have been proposed for interaction studies so far.

Historically, marginal analysis has prevailed in G×E interaction studies within the framework of genetic association studies. Although recent studies have confirmed the utility of regularized variable selection in joint G×E analysis, more efforts are needed for marginal penalizations, especially through the Bayesian point of view. The step toward marginal Bayesian variable selection is of particular significance in developing a coherent framework of analyzing G×E interactions.

Here, we propose a novel marginal Bayesian variable selection method for the robust identification of G×E interactions. As heavy-tailed distributions and outliers in the response variable have been widely observed, robust modeling is essential for yielding reliable results. Specifically, the robustness of the proposed method is facilitated by the Bayesian formulation of the least absolute deviation (LAD) regression, which has been a popular choice in frequentist G×E studies but seldom investigated in a similar context from the Bayesian perspective. We consider the Bayesian LAD LASSO for regularized identification of interaction effects. As Bayesian LAD LASSO does not lead to zero coefficients, the spike-and-slab priors (George and McCulloch, 1993; Ishwaran and Rao, 2005) has been incorporated to impose exact sparsity in the adaptive shrinkage framework. The corresponding MCMC algorithm has been developed to accommodate fast computations. We have demonstrated the advantage of the proposed robust Bayesian marginal analysis in simulation. The findings from the case study of the Nurses' Health Study (NHS) with SNP measurements have important biological implications.

2. METHOD

We use Y to denote a continuous response variable representing 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 X_{ij} $(j = 1, \ldots, p)$, the measurement of the jth genetic factor on the ith subject considers the following marginal model:

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

where α_k's and γ_t's are the regression coefficients corresponding to effects of environmental and clinical factors, respectively. For the jth gene X_j $(j = 1, \ldots, p)$, the G×E interactions effects are defined with $W_j = (X_j E_1, \ldots, X_j E_q)$, $\eta_j = (\eta_{j1}, \ldots, \eta_{jq})^T$. With a slight abuse of notation, denote $\tilde{W} = W_j$. The β_j's and η_{jk}'s are the regression coefficients of the genetic variants and G×E interactions effects, correspondingly. Let us denote $\alpha = (\alpha_1, \ldots, \alpha_q)^T$ and $\gamma = (\gamma_1, \ldots, \gamma_m)^T$. Then model (1) can be written as:

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

2.1. Bayesian Formulation of the LAD Regression

The necessity of accounting for robustness in interaction studies has been increasingly recognized (Zhou et al., 2021). Within the frequentist framework, it is essentially dependent on adopting a robust loss function to quantify lack of fit (Wu and Ma, 2015). Among a variety of popular robust losses, the least absolute deviation (LAD) loss function is well known for its advantages in dealing with heavy-tailed error distributions or outliers in response. The estimation of regression coefficients amounts to the following minimization problem:

$$\min_{\alpha,\gamma,\beta_j,\eta_j} \sum_{i=1}^{n} |Y_i - E_i \alpha - C_i \gamma - X_{ij} \beta_j - \tilde{W}_i \eta_j|.$$

Here, we propose the robust marginal Bayesian variable selection based on LAD. As the Laplace distribution is equivalent to the mixture of an exponential distribution and a scaled normal distribution (Kozumi and Kobayashi, 2011), for a Bayesian formulation of LAD regression, we assume that $\epsilon_i (i = 1, \ldots, n)$ are i.i.d. random variables following the Laplace distribution with density:

$$f(\epsilon_i|\tau) = \frac{\tau}{2} \exp(-\tau|\epsilon_i|),$$

where τ is the inverse of the scale parameters from the Laplace density. Then the likelihood function of our marginal G×E model can be expressed as:

$$f(Y|\alpha,\gamma,\beta_j,\eta_j)=\prod_{i=1}^{n}\frac{\tau}{2}\exp(-\tau|Y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j|).$$

The above formulation using Laplace distribution is a special case of the asymmetric Laplace distribution, which has been widely adopted in Baysian quantile regression (Yu and Moyeed, 2001; Yu and Zhang, 2005). In Baysian quantile regression, ϵ_i's are assumed to follow the skewed Laplace distribution with density

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

where $\rho_\theta(\epsilon)=\epsilon\{\theta-I(\epsilon<0)\}$ is the check loss function. The random errors can be written as

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

where

$$\xi_1=\frac{1-2\theta}{\theta(1-\theta)}\quad and\quad \xi_2=\sqrt{\frac{2}{\theta(1-\theta)}}$$

with quantile level $\theta\in(0,1)$, $\nu_i\sim\exp(\tau^{-1})$, and $z_i\sim N(0,1)$.

The Bayesian LAD regression is a special case of Bayesian quantile regression (Li et al., 2010) with θ=0.5, resulting in that $\xi_1=0$ and $\xi_2=\sqrt{8}$. Therefore, the response Y_i can be written as:

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

where $\mu_i=E_i\alpha+C_i\gamma+X_{ij}\beta_j+\tilde{W}_i\eta_j$.

2.2. Bayesian LAD LASSO With Spike-and-Slab Priors

In model (1), the coefficients β_j and η_j correspond to the main and interaction effects with respect to the jth genetic variant, respectively. When $\beta_j=0$ and $\eta_j=0$, the genetic variant has no effect on the phenotype. A non-zero β_j suggests the presence of main genetic effect. For η_j, if at least one of its component is not zero, then the G×E interaction effect exists. In literature, Bayesian quantile LASSO, with Bayesian LAD LASSO as its special case, has been proposed to conduct variable selection (Li et al., 2010). However, a major limitation is that Bayesian quantile LASSO cannot shrink regression coefficients to 0 exactly, resulting in inaccurate identification and biased estimation. To overcome such a limitation, we incorporate spike-and-slab priors to impose sparsity within Bayesian LAD LASSO framework as follows.

For the jth gene ($j=1,\ldots,p$), the marginal LAD LASSO model is given by

$$\sum_{i=1}^{n}|Y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j|+\lambda_1|\beta_j|+\lambda_2\sum_{k=1}^{q}|\eta_{jk}|.$$

Let $\varphi_1=\tau\lambda_1$ and $\varphi_2=\tau\lambda_2$. Then the conditional Laplace prior on the coefficient of main effect β_j can be expressed as scale mixtures of normals:

$$\begin{aligned}\pi(\beta_j|\tau,\lambda_1)&=\frac{\varphi_1}{2}\exp\{-\varphi_1|\beta_j|\}\\ &=\int_0^\infty\frac{1}{\sqrt{2\pi s_1}}\exp(-\frac{\beta_j^2}{2s_1})\frac{\varphi_1^2}{2}\exp(\frac{-\varphi_1^2}{2}s_1)ds_1.\end{aligned}$$

The conditional Laplace prior on the coefficients of interaction effect η_j can be written as:

$$\begin{aligned}\pi(\eta_j|\tau,\lambda_2)&=\prod_{k=1}^{q}\frac{\varphi_2}{2}\exp\{-\varphi_2|\eta_{jk}|\}\\ &=\prod_{k=1}^{q}\int_0^\infty\frac{1}{\sqrt{2\pi s_2}}\exp(-\frac{\eta_{jk}^2}{2s_2})\frac{\varphi_2^2}{2}\exp(\frac{-\varphi_2^2}{2}s_2)ds_2.\end{aligned}$$

Therefore, we consider the following hierarchical formulation for the marginal G×E model:

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

where $\delta_0(\beta_j)$ and $\delta_0(\eta_{jk})$ denote the spike at 0, respectively, and the slab distributions are represented by two normal distributions, $N(0,s_1)$ and $N(0,s_2k)$. Here, $\pi_1\in[0,1]$ and $\pi_2\in[0,1]$. The mixture of the spike and slab components facilitate the selection of main and interaction effects. Instead of setting π_1 and π_2 to a fixed value such as 0.5, we assign conjugate beta priors on them as $\pi_1\sim\text{Beta}(r_1,u_1)$ and $\pi_2\sim\text{Beta}(r_2,u_2)$, which account for the uncertainty in π_1 and π_2. In this paper, we choose $r_1=u_1=r_2=u_2=1$ as it gives a prior mean with 0.5 and it also allows a prior to spread out.

In addition, the normal prior has been placed on the coefficients of environmental factor $\alpha_k(k=1,\ldots,q)$ and clinical factor $\gamma_t(t=1,\ldots,m)$ 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 conjugate Gamma priors on τ, φ_1^2 and φ_2^2 with

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

In typical G×E studies, the environmental and clinical factors are of low dimensionality and the selection of them is not of interest.

Therefore, the sparsity-inducing priors have not been adopted for these factors. We consider the Bayesian LAD LASSO type of regularization in the proposed study as published studies have demonstrated that baseline penalty such as MCP and LASSO work well for marginal variable selection (Shi et al., 2014; Chai et al., 2017).

It is noted that Zhang et al. (2020) has proposed a marginal sparse group MCP to respect the strong hierarchy between main and interaction effects. Their results are promising when long tailed distributions and outliers are not present in the response variable. Although sparse group (or, bi-level) variable selection has been demonstrated as being very effective in multiple G×E studies based on joint models (Zhou et al., 2021), in our study, there is only one group per each marginal model. The sparse group no longer has significant advantages over individual level selection. Therefore, it has not been considered here.

Our model respects the weak hierarchy of "main effects, interactions." If imposing the strong hierarchy is needed, the genetic factor, once it is not selected given the presence of corresponding interaction effects, can be added back to the identified marginal model for a refit to impose strong hierarchy (Chai et al., 2017). While such a practice is not uncommon in marginal interaction studies, Shi et al. (2014) has also revealed satisfactory performance when strong hierarchy has not been pursued.

2.3. The Gibbs Sampler for Robust Marginal G×E Analysis

For the jth genetic factor, the joint posterior distribution of all the unknown parameters conditional on data can be expressed as

$$
\begin{aligned}
&\pi(\alpha,\gamma,\beta_j,\eta_j,\nu,s_1,s_2,\tau,\varphi_1,\varphi_2,\pi_1,\pi_2,z_i|Y)\\
&\propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2\nu_i}}\exp\Big\{-\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2\nu_i}\Big\}\\
&\times\prod_{i=1}^{n}\tau\exp(-\tau\nu_i)\tau^{a-1}\exp(-b\tau)\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z_i^2)\\
&\times\prod_{k=1}^{q}\frac{1}{\sqrt{(2\pi\alpha_0)}}\exp(-\frac{\alpha_k^2}{2\alpha_0})\\
&\times\prod_{t=1}^{m}\frac{1}{\sqrt{(2\pi\gamma_0)}}\exp(-\frac{\gamma_t^2}{2\gamma_0})\\
&\times\Big((1-\pi_1)(2\pi s_1)^{-1/2}\exp(-\frac{\beta_j^2}{2s_1})\mathbf{I}_{\{\beta_j\neq0\}}+\pi_1\delta_0(\beta_j)\Big)\\
&\times\prod_{k=1}^{q}\Big((1-\pi_2)(2\pi s_{2k})^{-1/2}\exp(-\frac{\eta_{jk}^2}{2s_{2k}})\mathbf{I}_{\{\eta_{jk}\neq0\}}+\pi_2\delta_0(\eta_{jk})\Big)\\
&\times\frac{\varphi_1^2}{2}\exp(-\frac{\varphi_1^2}{2}s_1)\\
&\times\prod_{k=1}^{q}\frac{\varphi_2^2}{2}\exp(-\frac{\varphi_2^2}{2}s_{2k})\\
&\times(\varphi_1^2)^{c_1-1}\exp(-d_1\varphi_1^2)\\
&\times(\varphi_2^2)^{c_2-1}\exp(-d_2\varphi_2^2)\\
&\times\pi_1^{r_1-1}(1-\pi_1)^{u_1-1}\\
&\times\pi_2^{r_2-1}(1-\pi_2)^{u_2-1}
\end{aligned}
$$

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

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

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

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

and variance

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

The posterior distribution of the coefficient of clinical factor $\gamma_t(t=1,\ldots,m)$ conditional on all other parameters can be obtained in similar way. Let $\mu_{(-\gamma_t)}=E(y_i)-C_{it}\gamma_t$, $i=1,\ldots,n$, then

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

where

$$
\begin{aligned}
\mu_{\gamma_t}&=(\sum_{i=1}^{n}\frac{\tau(y_i-\mu_{(-\gamma_t)})C_{it}}{\xi_2^2\nu_i})\sigma_{\gamma_t}^2,\\
\sigma_{\gamma_t}^2&=(\sum_{i=1}^{n}\frac{\tau C_{it}^2}{\xi_2^2\nu_i}+\frac{1}{\gamma_0})^{-1}.
\end{aligned}
$$

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 the coefficient of genetic factor β_j is a spike-and-slab distribution:

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

where

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

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

We can show that

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

The posterior distribution of β_j is a mixture of a normal distribution and a point mass at 0. That is, at each iteration 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.

Similarly, the posterior distribution of the interaction of the jth gene and environmental factors $\eta_{jk}(k = 1, \ldots, q)$ 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:

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

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

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

The full conditional posterior distribution of s_1 is:

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

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

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

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

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

Similarly, for $s_{2k}(k = 1, \ldots, q)$, when $\eta_{jk} = 0$, the posterior distribution of s_{2k} is $\exp(\frac{\varphi_2^2}{2})$. 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:

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

Therefore, the posterior distribution for φ_1^2 is Gamma ($c_1 + 1, s_1/2 + d_1$). Similarly, 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 is given as:

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

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 is given as:

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

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 is given as:

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

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

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

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

It is easy to show that

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

The spirit of marginal penalization for G×E interactions lies in the usage of a common sparsity cutoff to determine a list of important main and interaction effects. Instead of focusing on a fixed cutoff, varying the cutoff can generate different lists, resulting in a comprehensive view of important findings. The tuning parameter in penalized estimation serves as the cutoff. Therefore, the same tuning parameter has to be adopted for all the sub-models (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020). To further justify such a common tuning parameter, Zhang et al. (2020) has attempted using the joint model to select the common tuning through cross-validation. However, this seems not coherent with the nature of marginal analysis.

Ideally, the tuning parameter should be determined by each model itself to allow for flexibility in controlling sparsity individually, and a common cutoff is still available to examine different lists of important effects. With the Bayesian formulation, we can avoid such a limitation of frequentist marginal penalization methods. In particular, the priors have been placed on regularization parameters to determine the sparsity in a data-driven manner for each sub-model. With the spike-and-slab priors, the posterior distributions on the coefficients of main and interaction effects naturally lead to the usage of inclusion probability as a common cutoff to pin down the list of important effects, which is described in detail in the next section.

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, LAD Bayesian Lasso with spike-and-slab priors (denoted as LADBLSS) with three alternatives, LAD Bayesian Lasso (denoted as LADBL), Bayesian Lasso with spike-and-slab priors (denoted as BLSS) and Bayesian Lasso (denoted as BL). LADBL is similar to the proposed method, except that it does not adopt the spike-and-slab prior. The details of posterior inference are given in the **Appendix**.

Under all settings, the sample size is set as $n = 200$, and the number of G factors is $p = 500$ with $q = 4$, $m = 3$. For environmental factors, we simulate four continuous variables from multivariate normal distributions with marginal mean 0, marginal variance 1 and AR1 correlation structure with $\rho = 0.5$. In addition, three clinical factors are generated from a multivariate normal distribution with marginal mean 0 and marginal variance 1 and AR1 structure with $\rho = 0.5$. Among the p main G effects and pq G×E interactions, 8 and 12 effects are set as being associated with the response, respectively. All the environmental and clinical factors are important with nonzero coefficients, which are randomly generated 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 degrees 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.

In addition, the genetic factors are simulated in the following four settings.

Setting 1: 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 $\rho = 0.5$.

Setting 2: We assess the performance under single-nucleotide polymorphism (SNP) data. The SNPs are obtained by dichotomizing the gene expression values at the 1st and 3rd quartiles, with the 3 levels (0,1,2) for genotypes (aa, Aa, and AA). Here, the gene expressions are generated from the first setting.

Setting 3: Consider simulating the SNP data under a pairwise linkage disequilibrium (LD) structure. For the two minor alleles A and B of two adjacent SNPs, let q_1 and q_2 be the minor allele frequencies (MAFs). The frequencies of four haplotypes are as $p_{AB} = q_1q_2 + \delta$, $p_{ab} = (1 - q_1)(1 - q_2) + \delta$, $p_{Ab} = q_1(1 - q_2) - \delta$, and $p_{aB} = (1 - q_1)q_2 - \delta$, 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 $(q_1^2, 2q_1(1 - q_1), (1 - q_1)^2)$. Based on the conditional genotype probability matrix, we can simulate the genotypes for locus 2. With MAFs 0.3 and pairwise correlation $r = 0.6$, we have $\delta = r\sqrt{q_1(1 - q_1)q_2(1 - q_2)}$.

We collect the posterior samples from the Gibbs Sampler with 10,000 iterations and discard the first 5,000 samples as burn-ins. The posterior medians are used to estimate the coefficients. For approaches incorporating spike-and-slab priors, we consider computing the inclusion probability to indicate the importance of predictors. Here, we use a binary indicator ϕ to denote that the membership of the non-spike distribution. Take the main effect of the jth genetic factor, X_j, as an example. Suppose we have collected H posterior samples from MCMC after burn-ins. The *j*th G factor is included in the marginal G×E model at the *h*th MCMC iteration if the corresponding indicator is 1, i.e., $\phi_j^{(h)} = 1$. Subsequently, the posterior probability of retaining the *j*th genetic main effect in the final marginal model is defined as the average of all the indicators for the *j*th G factor among the H posterior samples. That is,

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

A larger posterior inclusion probability p_j indicates a stronger empirical evidence that the *j*th genetic main effect has a non-zero coefficient, i.e., a stronger association with the phenotypic trait.

To comprehensively assess the performance of the proposed and alternative methods, we consider a sequence of probabilities as cutting-offs in inclusion probability for methods with spike-and-slab priors. Given a cutoff probability, the main or interaction is included in the final marginal model if its posterior inclusion probability is larger than the cutoff, and is excluded otherwise. Provided with a sequence of cutting-off probabilities from small to large, we can investigate the set of identified effects and calculate the true/false positive rates (T/FPR) as the ground truth is known in simulation. For the sequence of cut-offs, we are able to compute the area under curve (AUC) as a comprehensive measure. Besides, for methods without spike-and-slab priors, the confidence level of the credible intervals can be adopted as the cut-off to compute TPR and FPRs. Therefore, all the methods under comparison can be evaluated on the same ground.

In addition, we also consider Top100, which is defined as the number of true signals when 100 important main effects (or interactions) are identified. For methods with spike-and-slab priors, 100 main effects or interactions are chosen with the highest inclusion probabilities. For methods without spike-and-slab priors, the indicators of all effects are computed for a sequence of credible levels. The top 100 main effects or interactions are chosen in terms of the highest average identification values.

Simulation results for the gene expression data in the first setting are tabulated in **Tables 1**, **2**. We can observe that the proposed method has the best performance among all approaches, especially when the response variable has heavy-tailed distributions. First, the performance of methods with spike-and-slab priors is consistently better than methods without spike-and-slab priors. For example, in **Table 1**, under error 3, the AUC of LADBLSS is 0.9558 (sd 0.0161), which is much larger than that of the robust method without spike-and-slab priors, i.e., 0.8432(sd 0.0115) from LADBL. Also, the AUC of robust methods is much larger than that of non-robust methods, especially in the presence of heavy-tailed errors. For instance, in the first setting under error 3, the AUC of LADBLSS is 0.9558 and the AUC of LADBL is 0.8432 while that of BLSS and BL is around 0.5. Similar advantageous performance can also be observed from the identification results with Top100. In **Table 2** under error 5, LADBLSS identifies 7.80 (sd 0.55) out of the 8 main effects and 10.53 (sd 1.36) out of the 12 interaction effects. This is higher than the results of LADBL with 7.57 (sd 0.57) of main effects and 6.83 (sd 1.07) of interaction effects. Second, among all the methods with spike-and-slab priors, Bayesian LAD method with spike-and-slab priors has the best performance in all identification results. Under error 3, in **Table 1**, the AUC of LADBLSS is 0.9558(sd 0.0161) while the AUC of BLSS is 0.5473(sd 0.0576). Under error 4 in **Table 2**, LADBLSS identifies 7.77(sd 0.57) main effects and 10.67(sd 1.50) interaction effects while BLSS identifies 6.2(sd 2.62) main effects and 8.3(sd 3.98) interaction effects, respectively.

TABLE 1 | Simulation results of the first setting for BL (Bayesian LASSO), BLSS (Bayesian LASSO with spike-and-slab priors), LADBL (LAD Bayesian LASSO), and LADBLSS (LAD Bayesian LASSO with spike-and-slab priors).

		BL	BLSS	LADBL	LADBLSS
Error 1	AUC	0.9182	0.9901	0.9258	0.9887
N(0,1)	SD	0.0052	0.0021	0.0076	0.0026
Error 2	AUC	0.8332	0.9420	0.9004	0.9841
t(2)	SD	0.0107	0.0235	0.0078	0.0031
Error 3	AUC	0.5343	0.5473	0.8432	0.9558
Lognormal(0,2)	SD	0.0144	0.0576	0.0115	0.0161
Error 4	AUC	0.8221	0.9124	0.9222	0.9895
90%N(0,1) + 10%Cauchy(0,1)	SD	0.0212	0.0410	0.0071	0.0024
Error 5	AUC	0.7507	0.8431	0.9192	0.9904
80%N(0,1) + 20%Cauchy(0,1)	SD	0.0217	0.0633	0.0059	0.0018

AUC (mean of AUC) and SD (sd of AUC) based on 100 replicates. n = 200, p = 500, q = 4, and m = 3.

Similar patterns can be observed in Tables 4, 5 in **Appendix** for the second setting, and Tables 6, 7 in **Appendix** for the third setting in **Appendix**. We have also investigated the performance of when n = 2,000 under setting 1. While the difference among the 4 methods significantly diminishes with such a large sample size, we can still observe the superior performance of LADBLSS by using a shorter list of top ranked effects. The results are provided in the table from **Supplementary Material**. Overall, the advantages of conducting robust Bayesian G×E analysis using the proposed approach can be justified based on the results of comprehensive simulation studies. The convergence of the MCMC chains with the potential scale reduction factor (PSRF) (Brooks and Gelman, 1998) has been conducted. In this study, we use PSRF $\leq$ 1.1 (Gelman et al., 2004) as the cut-off point, which indicates that chains converge to a stationary distribution. The convergence of chains after burn-ins has been checked for all parameters with the value of PSRF <1.1. **Figure 1** shows the convergence pattern of PSRF for the main and interaction coefficients of the first genetic factors in Example 1 under error 3.

In simulation, the hyperparameters for the Gamma priors and Beta priors specified in section Bayesian LAD LASSO With Spike-and-slab Priors are set to 1. In addition, the initial values of the regression parameters are also set to 1. Based on our experiments, the results and convergence of the MCMC algorithm are not sensitive to the choice of these parameters. We have observed satisfactory convergence for all of our simulations. For one simulated dataset under the first setting with n = 200, p = 500

TABLE 2 | Identification results of the first setting with Top100 method for BL (Bayesian LASSO), BLSS (Bayesian LASSO with spike-and-slab priors), LADBL (LAD Bayesian LASSO) and LADBLSS (LAD Bayesian LASSO with spike-and-slab priors).

		Main	Interaction	Total
Error 1	BL	7.60(0.49)	6.80(1.6)	14.40(1.73)
N(0,1)	BLSS	7.80(0.41)	10.80(0.92)	18.60(1.13)
	LADBL	7.67(0.55)	6.53(1.85)	14.20(1.81)
	LADBLSS	7.76(0.5)	10.53(1.36)	18.30(1.49)
Error 2	BL	6.37(1.90)	3.90(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)	5.80(1.71)	13.23(2.01)
	LADBLSS	7.53(0.51)	9.90(1.56)	17.43(1.76)
Error 3	BL	0.90(1.21)	0.50(0.97)	1.40(1.45)
Lognormal(0,2)	BLSS	0.73(0.94)	0.47(0.68)	1.20(1.35)
	LADBL	6.27(1.55)	3.67(1.94)	9.93(2.75)
	LADBLSS	6.10(1.37)	8.93(2.02)	15.03(3.09)
Error 4	BL	5.57(2.99)	3.63(2.53)	9.20(5.05)
90%N(0,1)	BLSS	6.20(2.62)	8.30(3.98)	14.50(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)	3.00(2.49)	8.07(5.01)
80%N(0,1)	BLSS	4.60(3.25)	5.70(4.23)	10.30(7.27)
+20%Cauchy(0,1)	LADBL	7.57(0.57)	6.83(1.07)	14.40(1.83)
	LADBLSS	7.80(0.55)	10.53(1.36)	18.33(1.69)

Mean(sd) based on 100 replicates. $n = 200$, $p = 500$, $q = 4$, and $m = 3$.

and standard normal error, the CPU time (in minutes) for fitting all the 500 marginal models through 10,000 MCMC iterations on a laptop with standard configurations are 1.27(BL), 1.75(BLSS), 6.16(LADBL), and 5.95 (LADBLSS) minutes, respectively. The source codes of implementing all the methods under comparison are included in the **Supplementary Material**.

4. REAL DATA ANALYSIS

In this study, we analyze the type 2 diabetes (T2D) data from Nurses' Health Study (NHS), which is a well-characterized cohort study of women with high-dimensional SNP data, as well as measurements on lifestyle and dietary factors. We consider SNPs on chromosome 10 to identify main and gene–environment interactions associated with weight, which is an important phenotypic trait related to type 2 diabetes. Here, weight is used as response and five environment factors, age (age), total physical activity (act), trans fat intake (trans), cereal fiber intake (ceraf), and reported high blood cholesterol (chol), are considered. Data are available on 3,391 subjects and 17,016 gene expressions after cleaning the raw data through matching phenotypes and genotypes and removing SNPs with MAF <0.05. A prescreening is done before downstream analysis. We use a marginal linear model with weight as response and age, act, trans, ceraf, and chol as environment factors. Note that 10,000 SNPs that have at least two main or interaction effects with $p <$ 0.05 are kept.The scale of working data is generally not a major concern for marginal analysis, as the computation can be done in a highly parallel manner. Here, we focus on chromosome 10 which has been reported to harbor interesting genes in existing studies.

We use Top 100 method to identify 100 most important main and interaction effects. The proposed method LADBLSS identifies 20 main SNP effects and 80 gene–environment interactions, which are listed in Table 8 in **Appendix**. Our study provides crucial implications in identifying the important main and interactions of SNPs and its associations with weight. For example, three SNPs, rs17011106, rs4838643 and rs17011115, located within gene WDFY4 are identified. WDFY4 has been observed as an influential factor related to weight and obesity (Barclay et al., 2015; Martin et al., 2019). In addition, SNPs rs10994364, rs10821773, and rs10994308, located within gene ANK3, are identified with interacting environment factors age and chol. There are findings showing an association between ANK3 and higher systolic blood pressure (Ghanbari et al., 2014). Published studies have also shown that ANK3 is linked to pulmonary and renal hypertension (Ghanbari et al., 2014). Allele risk variants have been identified in ANK3, and these variants explain a proportion of the heritability of BD (bipolar disorder), which is associated with higher body mass index (BMI) and increased metabolic comorbidity and the genetic risk for BD relates to common genetic risk with T2D (Winham et al., 2014). Our proposed method identifies its interaction with chol, the high blood cholesterol. Data from several sources suggest that islet cholesterol metabolism contributes to the pathogenesis of T2D (Brunham et al., 2008). Furthermore, the SNP rs1244416, corresponding to gene ATP5C1, interacts with the reported high blood cholesterol. This gene has been found to be deregulated in T2D skeletal muscle through pathway-based microanalysis (Morrison et al., 2012). The interactions between SNP rs10857590 and trans fat intake has also been identified by using the proposed method. The SNP is within gene ARHGAP22, which has been investigated in Huang et ail. (2018). As a diabetic retinopathy (DR) susceptibility gene, the expression of ARHGAP22 is positively associated with endothelial progenitor cells (EPC) levels in T2D patients with DR.

Analysis with alternatives BL, BLSS, and LADBL has also been conducted. To compare the alternative methods with the proposed method, we provide the numbers of main effects and interactions identified by these methods with pairwise overlaps in **Table 3**. It clearly shows that the proposed one results in a very different set of effects compared to alternatives. We refit the regularized marginal models by LADBL and LADBLSS using robust Bayesian Lasso, and those identified by BL and BLSS using Bayesian Lasso. In addition, the inclusion probabilities of the selected main and interaction effects using LADBLSS are provided in Table 9 in **Appendix**. Results from the alternative methods are available from the **Supplementary Material**. The proposed method selects the 100 most important effects with the inclusion probability larger than 0.9, demonstrating its superiority in quantifying uncertain compared to marginal penalization methods (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020). We noticed

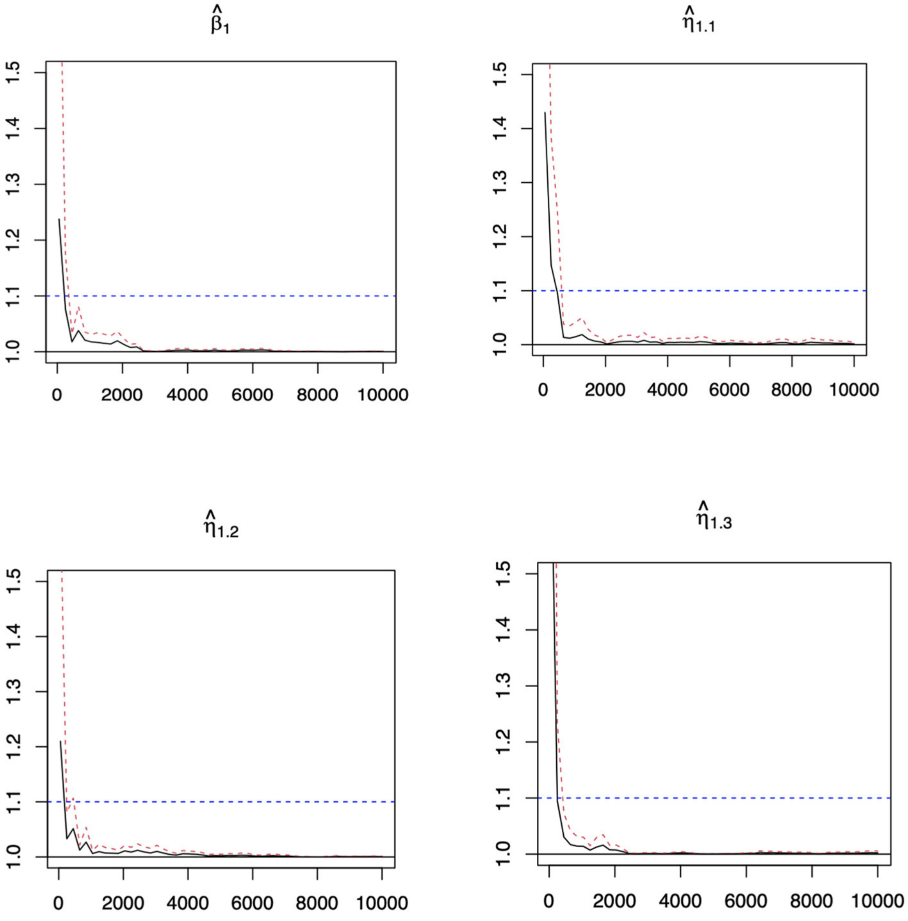

FIGURE 1 | Potential scale reduction factor (PSRF) against iterations for the coefficients of the first genetic factors and its interaction with environmental factors in Example 1 under error 3. Black line: the PSRF. Red dotted line: the upper limits of the 95% confidence interval for the PSRF. Blue dotted line: The threshold of 1.1. The $\hat{\beta}_1$ represents the estimated coefficients of the main effects for the first genetic factor. The $\hat{\eta}_{11}$ to $\hat{\eta}_{13}$ represent the estimated coefficients of the first three interaction effects for the first genetic factor.

the small magnitude of refitted regression coefficients from LAD-based methods compared to those obtained by the non-robust method in the **Supplementary Material**. This is due to the difference between the LAD-based and least square based loss function for robust and non-robust methods, respectively. The advantage of LADBLSS over the non-robust methods can be clearly observed. First, majority of the top 100 important effects identified by BL are main genetic effects. This is less likely to be reasonable as the response variable weight has been well acknowledged to be also dependent on gene–environment interactions. For BLSS, the inclusion probabilities are low compared to those of the LADBLSS, suggesting lower level of certainty and confidence in the regression coefficients obtained from BLSS. The inferior performance of BL and BLSS further justifies the need of developing robust methods in marginal gene–environment interaction studies. Overall, LADBLSS leads to identification results significantly different from all the alternatives, as well as main and interaction effects of important biological implications that are not discovered by the benchmarks.

TABLE 3 | The numbers of main G effects and interactions identified by different approaches and their overlaps for BL (Bayesian LASSO), BLSS (Bayesian LASSO with spike-and-slab priors), LADBL (LAD Bayesian LASSO), and LADBLSS (LAD Bayesian LASSO with spike-and-slab priors).

T2D	Main				Interaction			
	BL	BLSS	LADBL	LADBLSS	BL	BLSS	LADBL	LADBLSS
BL	86	5	6	8	14	14	4	8
BLSS		24	3	6		76	20	23
LADBL			20	12			80	50
LADBLSS				20				80

5. DISCUSSION

In the past, G×E interaction studies have been mainly conducted through marginal hypothesis testing, based on a diversity of study designs utilizing parametric, nonparametric, and semiparametric models (Murcray et al., 2009; Thomas, 2010; Mukherjee et al., 2012), which later have been extended to joint analyses driven primarily by the pathway or gene set based association studies (Wu and Cui, 2013a; Jin et al., 2014; Jiang et al., 2017). In addition, published literature has also reported the success of marginal screening studies, including those based on partial correlations (Niu et al., 2018; Xu et al., 2019). Recently, the effectiveness of regularized variable selection in G×E interaction studies has been increasingly recognized, and a large number of regularization methods have been proposed for joint interaction studies (Zhou et al., 2021). Marginal penalization has also been demonstrated as promising competitors, although they have only been investigated in a limited number of frequentist studies (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020).

Therefore, the proposed marginal robust Bayesian variable selection is of particular importance, since joint and marginal analysis cannot replace each other and marginal Bayesian penalization has not been examined for G×E studies so far. In particular, with the robustness and incorporation of spike-and-slab priors in the adaptive Bayesian shrinkage, the LADBLSS has an analysis framework more coherent with that of the joint robust analysis[1], which significantly facilitates methodological developments for interaction studies.

Nevertheless, the proposed method has limitations. As a fully Bayesian methods based on MCMC algorithms, the computation cost is generally high due to the tradeoff for quantifying uncertainty using posterior samples. Such a drawback can be addressed through conducting the computation in a parallel manner given the marginal nature of the method. Besides, the variable selection conducted in our study is based on the L1 penalty within the Bayesian framework. As this structure ignores the correlation among genetic features, a possible direction for future improvement is to incorporate network or gene set information in the identification of important gene–environment interactions (Wang et al., 2021). Furthermore, in our study, the genetic factor is represented by one SNP coded as a triadic factor. A closer look at both the additive and dominant penetrance effects of the SNP will lead to elucidation of the genetic basis using marginal interaction studies on a finer scale. For gene–environment interaction studies, marginal and joint analysis are the two major paradigms, and cannot replace each other (Zhou et al., 2021). It is always on a safe side to perform marginal analysis in G×E studies in addition to the joint ones, facilitating a more comprehensive understanding on the genetic architecture of complex diseases.

The marginal Bayesian regularization can be extended to different types of response, for example, under binary, categorical, prognostic and multivariate outcomes. Nevertheless, considering robustness in the generalized models with the Bayesian framework is not trivial, especially under the multivariate responses (Wu et al., 2014; Zhou et al., 2019). We postpone the investigations to the future studies.The interaction between genetic and environmental factors in this study has been modeled as the product of the two corresponding variables, which amounts to "linear" interactions. In practice, the linear interaction assumption has been frequently violated (Ma et al., 2011; Wu and Cui, 2013b; Zhao et al., 2019), which demands accommodation of these nonlinear effects through nonparametric and semiparametric models (Li et al., 2015; Wu et al., 2015, 2018; Ren et al., 2020). It is of great interest and importance to migrate the nonlinear G×E studies to marginal cases in the near future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Authorized access should be granted before accessing the data. Applications to access the data should be sent to dbGap (accession number phs000091.v2.p1). For more information, please refer to NIH dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by This study is a secondary data analysis. The dataset has been applied through NIH dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id =phs000091.v2.p1). In the dataset, the patient information has been de-identified. As indicated from the dbGap website under section Authorized Access/Use Restrictions, IRB is not required for accessing and using the data. According to the original publication, The study was approved by the institutional review board of Brigham and Women's Hospital in Boston; completion of the self-administered questionnaire was considered to imply informed consent. For more information regarding study population, please refer to the original publication: Hu et al. (2001). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL and CW: conceptualization and writing—original draft preparation. XL, KF, JR, and CW: methodology and writing—review and editing. XL: data analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank the editors and reviewers for their careful review and insightful comments. This work was supported by an innovative research award from K State Johnson Cancer Research Center.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.667074/full#supplementary-material

[1] Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y., and Wu, C. (under revision). Robust Bayesian variable selection for gene-environment interactions. *Biometrics*.

REFERENCES

Barclay, S. F., Rand, C. M., Borch, L. A., Nguyen, L., Gray, P. A., Gibson, W. T., et al. (2015). Rapid-Onset Obesity with Hypothalamic Dysfunction, Hypoventilation, and Autonomic Dysregulation (ROHHAD): exome sequencing of trios, monozygotic twins and tumours. *Orphanet J. Rare Dis.* 10:103. doi: 10.1186/s13023-015-0314-x

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Brunham, L. R., Kruit, J. K., Verchere, C. B., and Hayden, M. R. (2008). Cholesterol in islet dysfunction and type 2 diabetes. *J. Clin. Invest.* 118, 403–408. doi: 10.1172/JCI33296

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S. E., et al. (2017). Identifying gene-environment interactions for prognosis using a robust approach. *Econometr. Stat.* 4, 105–120. doi: 10.1016/j.ecosta.2016.10.004

Cordell, H. J., and Clayton, D. G. (2005). Genetic association studies. *Lancet* 366, 1121–1131. doi: 10.1016/S0140-6736(05)67424-7

Cornelis, M. C., and Hu, F. B. (2012). Gene-environment interactions in the development of type 2 diabetes: recent progress and continuing challenges. *Ann. Rev. Nutr.* 32, 245–259. doi: 10.1146/annurev-nutr-071811-150648

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004). *Bayesian Data Analysis*. London; Boca Raton, FL: Chapman and Hall/CRC.

George, E. I., and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* 88, 881–889. doi: 10.1080/01621459.1993.10476353

Ghanbari, M., de Vries, P. S., de Looper, H., Peters, M. J., Schurmann, C., Yaghootkar, H., et al. (2014). A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. *Hum. Mutat.* 35, 1524–1531. doi: 10.1002/humu.22706

Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., et al. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N. Engl. J. Med.* 345, 790–797. doi: 10.1056/NEJMoa010492

Huang, J., and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* 16, 176–195. doi: 10.1007/s10985-009-9144-2

Huang, Y. C., Liao, W. L., Lin, J. M., Chen, C. C., Liu, S. P., Chen, S. Y., et al. (2018). High levels of circulating endothelial progenitor cells in patients with diabetic retinopathy are positively associated with ARHGAP22 expression. *Oncotarget* 9, 17858. doi: 10.18632/oncotarget.24909

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nat. Rev. Genet.* 6, 287–298. doi: 10.1038/nrg1578

Ishwaran, H., and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* 33, 730–773. doi: 10.1214/009053604000001147

Jiang, Y., Huang, Y., Du, Y., Zhao, Y., Ren, J., Ma, S., et al. (2017). Identification of prognostic genes and pathways in lung adenocarcinoma using a Bayesian approach. *Cancer Inform.* 1:1176935116684825. doi: 10.1177/1176935116684825

Jin, L., Zuo, X., Su, W., Zhao, X., Yuan, M., Han, L., et al. (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics* 12, 210–220. doi: 10.1016/j.gpb.2014.10.002

Kozumi, H., and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *J. Stat. Comput. Simul.* 81, 1565–1578. doi: 10.1080/00949655.2010.496117

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. *Ann. Appl. Stat.* 9, 640. doi: 10.1214/15-AOAS808

Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Anal.* 5, 533–556. doi: 10.1214/10-BA521

Liu, C., Ma, J., and Amos, C. I. (2015). Bayesian variable selection for hierarchical gene–environment and gene–gene interactions. *Hum. Genet.* 134, 23–36. doi: 10.1007/s00439-014-1478-5

Ma, S., Yang, L., Romero, R., and Cui, Y. (2011). Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics* 27, 2119–2126. doi: 10.1093/bioinformatics/btr318

Martin, C. L., Jima, D., Sharp, G. C., McCullough, L. E., Park, S. S., Gowdy, K. M., et al. (2019). Maternal pre-pregnancy obesity, offspring cord blood DNA methylation, and offspring cardiometabolic health in early childhood: an epigenome-wide association study. *Epigenetics* 4, 325–340. doi: 10.1080/15592294.2019.1581594

Morrison, F., Johnstone, K., Murray, A., Locke, J., and Harries, L. W. (2012). Oxidative metabolism genes are not responsive to oxidative stress in rodent beta cell lines. *Exp. Diabetes Res.* 2012:793783. doi: 10.1155/2012/793783

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.* 175, 177–190. doi: 10.1093/aje/kwr367

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* 169, 219–226. doi: 10.1093/aje/kwn353

Niu, Y. S., Hao, N., and Zhang, H. H. (2018). Interaction screening by partial correlation. *Stat Its Interface* 11, 317–325. doi: 10.4310/SII.2018.v11.n2.a9

Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., et al. (2020). Semiparametric Bayesian variable selection for gene-environment interactions. *Stat. Med.* 39, 617–638. doi: 10.1002/sim.8434

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y., and Ma, S. (2014). A penalized robust method for identifying gene–environment interactions. *Genet. Epidemiol.* 38, 220–230. doi: 10.1002/gepi.21795

Simonds, N. I., Ghazarian, A. A., Pimentel, C. B., Schully, S. D., Ellison, G. L., Gillanders, E. M., et al. (2016). Review of the gene-environment interaction literature in cancer: what do we know?. *Genetic Epidemiol.* 40, 356–365. doi: 10.1002/gepi.21967

Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Ann. Rev. Public Health* 31, 21–36. doi: 10.1146/annurev.publhealth.012809.103619

Von Mutius, E. (2009). Gene-environment interactions in asthma. *J. Allergy Clin. Immunol.* 123, 3–11. doi: 10.1016/j.jaci.2008.10.046

Wang, H., Ye, M., Fu, Y., Dong, A., Zhang, M., Feng, L., et al. (2021). Modeling genome-wide by environment interactions through omnigenic interactome networks. *Cell Rep.* 35, 109114. doi: 10.1016/j.celrep.2021.109114

Winham, S. J., Cuellar-Barboza, A. B., Oliveros, A., McElroy, S. L., Crow, S., Colby, C., et al. (2014). Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. *Mol. Psychiatry* 19, 1010–1016. doi: 10.1038/mp.2013.159

Wu, C., and Cui, Y. (2013a). Boosting signals in gene-based association studies via efficient SNP selection. *Brief. Bioinformatics* 15, 279–291. doi: 10.1093/bib/bbs087

Wu, C., and Cui, Y. (2013b). A novel method for identifying nonlinear gene–environment interactions in case–control association studies. *Hum. Genet.* 132, 1413–1425. doi: 10.1007/s00439-013-1350-z

Wu, C., Cui, Y., and Ma, S. (2014). Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. *Stat. Med.* 33, 4988–4998. doi: 10.1002/sim.6287

Wu, C., Li, S., and Cui, Y. (2012). Genetic association studies: an information content perspective. *Curr. Genomics* 13, 566–573. doi: 10.2174/138920212803251382

Wu, C., and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics, *Brief. Bioinformatics* 16, 873–883. doi: 10.1093/bib/bbu046

Wu, C., Shi, X., Cui, Y., and Ma, S. (2015). A penalized robust semiparametric approach for gene–environment interactions. *Stat. Med.* 34, 4016–4030. doi: 10.1002/sim.6609

Wu, C., Zhong, P. S., and Cui, Y. (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Stat. Appl. Genet. Mol. Biol.* 17:/j/sagmb.2018.17.issue-2/sagmb-2017-0008/sagmb-2017-0008.xml. doi: 10.1515/sagmb-2017-0008

Xu, Y., Wu, M., Zhang, Q., and Ma, S. (2019). Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics* 111, 1115–1123. doi: 10.1016/j.ygeno.2018.07.006

Yu, K., and Moyeed, R. A. (2001). Bayesian quantile regression. *Stat. Probab. Lett.* 54, 437–447. doi: 10.1016/S0167-7152(01)00124-9

Yu, K., and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Commun. Stat. Theory Methods* 34, 1867–1879. doi: 10.1080/03610920500199018

Zhang, S., Xue, Y., Zhang, Q., Ma, C., Wu, M., and Ma, S. (2020). Identification of gene–environment interactions with marginal penalization. *Genet. Epidemiol.* 44, 159–196. doi: 10.1002/gepi.22270

Zhao, N., Zhang, H., Clark, J. J., Maity, A., and Wu, M. C. (2019). Composite kernel machine regression based on likelihood ratio test for joint testing of genetic and gene–environment interaction effect. *Biometrics* 75, 625–637. doi: 10.1111/biom.13003

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. doi: 10.3390/genes10121002

Zhou, F., Ren, J., Lu, X., Ma, S., and Wu, C. (2021). Gene–Environment Interaction: a Variable Selection Perspective. Epistasis. *Methods Mol. Biol.* 2212, 191–223. doi: 10.1007/978-1-0716-0947-7_13

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.