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

Xi Lu
(joint author: Kun Fan, Cen Wu)

Department of Statistics
Kansas State University

August 4, 2020

Background

- Many studies have demonstrated the advantages of variables selection methods in detecting $G \times E$ interactions from the frequentist point of view.
  - Marginal methods: Cornelis et al., 2011; Ma et al., 2011; Wu and Cui, 2013.
  - Joint methods: Liu et al., 2013; Shi et al., 2014; Wu et al., 2014; Wu et al., 2015; Wu et al., 2018.
- Bayesian variable selection, however, has not been widely developed for interaction studies.
- Existing Bayesian variable selection methods for $G \times E$ interaction
  - Linear interaction only: Oh et al., 2003 (SSVS); Ahn et al., 2013 (SSVS); Liu et al., 2013 (SSVS);
  - Non-linear interaction: Li et al., 2015 (Laplacian shrinkage).

Motivation

- Outliers and data contamination in disease phenotypes of $G \times E$ studies have been commonly encountered.
- Non-robust variable selection approaches can lead to biased estimates and false identifications.
- Robust Bayesian methods have not been investigated for gene-environment interactions by far.

The Margianl $G \times E$ Model

- For the jth gene X_j $(j = 1, \ldots, p)$, consider the marginal model:

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

Denote the ith subject with i. Let $X = (X_1, \ldots, X_p)$ be p genetic variants, $E = (E_1, \ldots, E_q)$ be q environmental factors, $C = (C_1, \ldots, C_m)$ be m clinical factor, $W_j = (X_j E_1, \ldots, X_j E_q)$, $\eta_j = (\eta_{j1}, \ldots, \eta_{jq})^T$.

The Margianl $G \times E$ Model

- Model (1) can be written as

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

Denote $\alpha = (\alpha_1, \ldots, \alpha_q)^T$, $\gamma = (\gamma_1, \ldots, \gamma_m)^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$, $\eta = (\eta_1^T, \ldots, \eta_p^T)^T$.

The LAD regression

- The least absolute deviation (LAD) regerssion is well known for its robustness to heavy-tailed errors or outliers in response.
- The regression coefficients can be estimated as the solution to the following minimization problem

$$\min_{\alpha,\gamma,\beta_j,\eta_j} \sum_{i=1}^{n} |Y_i - E_i\alpha - C_i\gamma - X_{ij}\beta_j - W_j\eta_j|. \tag{3}$$

Bayesian quantile regression

- In Baysian quantile regression, we assume that $\epsilon_i (i = 1, \ldots, n)$ following the skewed Laplace distribution with density:

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

- The random errors can be written as

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

- where

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

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

Bayesian formulation of the LAD regression

- Bayesian LAD regression is a special case of Bayesian quantile regression with θ=0.5.
- We have $\xi_1 = 0$ and $\xi_2 = \sqrt{8}$.
- The Bayesian LAD regression model can be written as:

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

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

Issues in analyzing genomic data for $G \times E$ interactions

- The high-dimensionality of genomic data.
- Not all genetic factors have interactions with the environment factors.

Bayesian LAD regression with spike-and-slab priors

- A non-zero β suggests a presence of main effect and a non-zero η suggests an interaction effect.
- We incorporate spike-and-slab priors to impose sparsity.
- For the robust Bayesian marginal model of the jth gene ($j = 1, \ldots, p$), consider the following priors:

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

Bayesian Hierarchical Structure

- π_1 and π_2 control the sparisity on the main and interaction level.
- Assign $\pi_1 \sim \text{Beta}(r_1, u_1)$ and $\pi_2 \sim \text{Beta}(r_2, u_2)$.
- Place normal priors on $\alpha_k (k = 1, \ldots, q)$ and $\gamma_t (t = 1, \ldots, m)$ as:

$$\alpha_k \overset{iid}{\sim} \frac{1}{\sqrt{(2\pi\alpha_0)}} \exp(-\frac{\alpha_k^2}{2\alpha_0}), (k = 1, \ldots, q),$$

$$\gamma_t \overset{iid}{\sim} \frac{1}{\sqrt{(2\pi\gamma_0)}} \exp(-\frac{\gamma_t^2}{2\gamma_0}), (t = 1, \ldots, m).$$

- Assume Gamma priors on τ, φ_1^2 and φ_2^2 with

$$\tau \sim \text{Gamma}(a, b),$$

$$\varphi_1^2 \sim \text{Gamma}(c_1, d_1),$$

$$\varphi_2^2 \sim \text{Gamma}(c_2, d_2).$$

The Gibbs Sampler

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

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

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

- The full conditional posterior distribution of γ_t is:

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

with

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

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

The Gibbs Sampler

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

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

where

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

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

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

The Gibbs Sampler

- Denote $\tilde{W} = W_j$, let $\mu_{(-\eta_{jk})} = E(Y_i) - \tilde{W}_{ik}\eta_{jk}$ and $l_{2k} = \pi(\eta_{jk} = 0|\text{rest})$, η_{jk} follows this distribution:

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

where

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

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

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

The Gibbs Sampler

- When $\beta_j = 0$, $s_1|\text{rest} \sim \exp(\frac{\varphi_1^2}{2})$.

  When $\beta_j \neq 0$, $s_1^{-1}|\text{rest} \sim \text{Inverse-Gaussian}(\sqrt{\frac{\varphi_1^2}{\beta_j^2}}, \varphi_1^2)$.

- When $\eta_{jk} = 0$, $s_{2k}|\text{rest} \sim \exp(\frac{\varphi_2^2}{2})$.

  When $\eta_{jk} \neq 0$, $s_{2k}^{-1}|\text{rest} \sim \text{Inverse-Gaussian}(\sqrt{\frac{\varphi_2^2}{\eta_{jk}^2}}, \varphi_2^2)$.

- $\varphi_1^2|\text{rest} \sim \text{Gamma}(c_1 + 1, s_1/2 + d_1)$

  $\varphi_2^2|\text{rest} \sim \text{Gamma}(c_2 + q, \sum_{k=1}^{q} s_{2k}/2 + d_2)$

- $\pi_1|\text{rest} \sim \text{Beta}(1 + r_1 - \mathbf{I}(\beta_j \neq 0), u_1 + \mathbf{I}(\beta_j \neq 0))$

  $\pi_2|\text{rest} \sim \text{Beta}(1 + r_1 - \sum_{k=1}^{q} \mathbf{I}(\eta_{jk} \neq 0), u_1 + \sum_{k=1}^{q} \mathbf{I}(\eta_{jk} \neq 0))$

- $\tau|\text{rest} \sim \text{Gamma}(a + \frac{3}{2}n, \ [\sum_{i=1}^{n}(\frac{(y_i - E_i\alpha - C_i\gamma - X_{ij}\beta_j - \tilde{W}_i\eta_j)^2}{2\xi_2^2 v_i} + v_i) + b])$

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

- $j = 1, \ldots, p; \ k = 1, \ldots, q.$

Simulation

- n=200, p=500, q=4, m=3.
- X is generated from multivariate normal distribution with marginal mean 0 and variance 1 with an AR structure with $\rho = 0.5$.
- C is generated from multivariate normal distribution with marginal mean 0 and variance 1.
- case1: E is a continuous variable generated from multivariate normal distribution.
- case2: E is a mixture of continuous variables and discrete variables.
- Among the p main G effects and $p \times q$ $G \times E$ interactions, 8 and 12 are set as associated with the response.
- All environmental factors and clinical factors have important effects.

Simulation

- Nonzero coefficients 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) $t(2)$(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).
- Four approaches:
  - BL: Bayesian LASSO;
  - BLSS: Bayesian LASSO with Spike-and-Slab priors;
  - LADBL: Robust Bayesian LASSO;
  - LADBLSS: Robust Bayesian LASSO with Spike-and-Slab priors.

Simulation

- Posterior samples are collected from a Gibbs Sampling with 10,000 iterations in which the first 5,000 are burn-ins.
- We use posterior median to estimate the coefficients.
- For methods incorporating spike-and-slab priors, we consider the median probability model (MPM) to identify significant predictors.
- For methods without spike-and-slab priors, the 95% credible interval is used.
- To compare these methods, we use ROC curves, AUC and Top100.

ROC curves

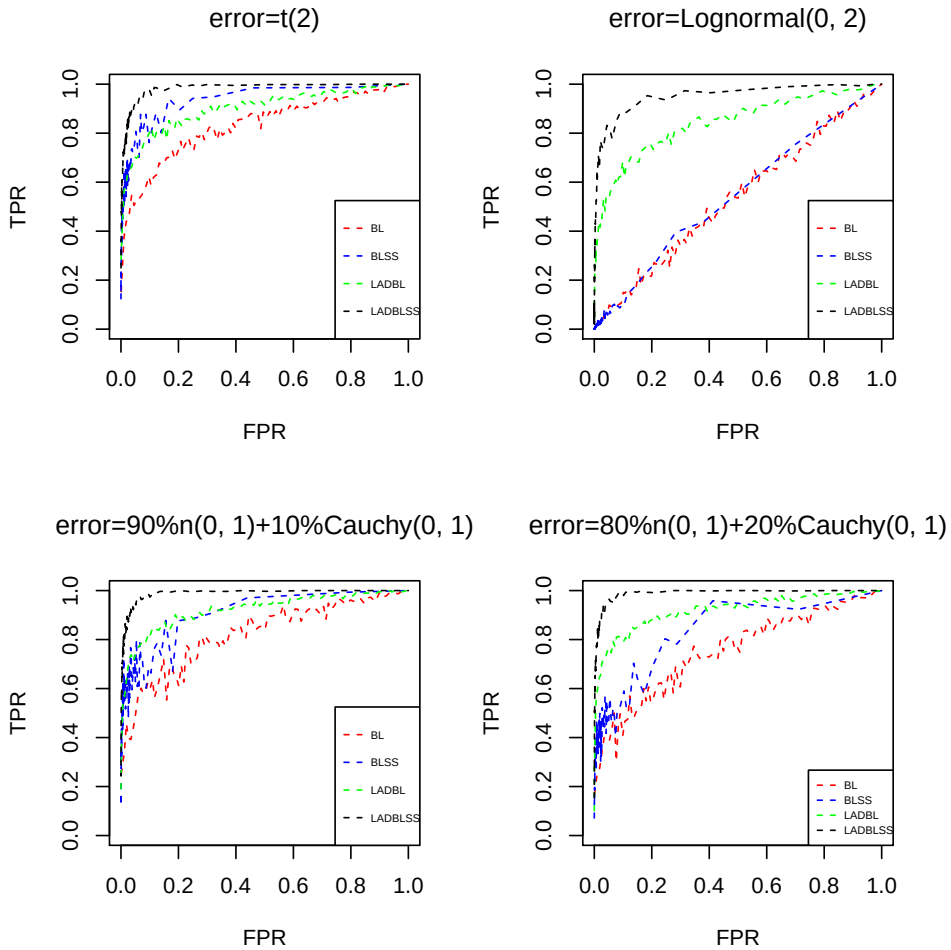

Figure: ROC curves of case1

Identification results

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

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

ROC curves

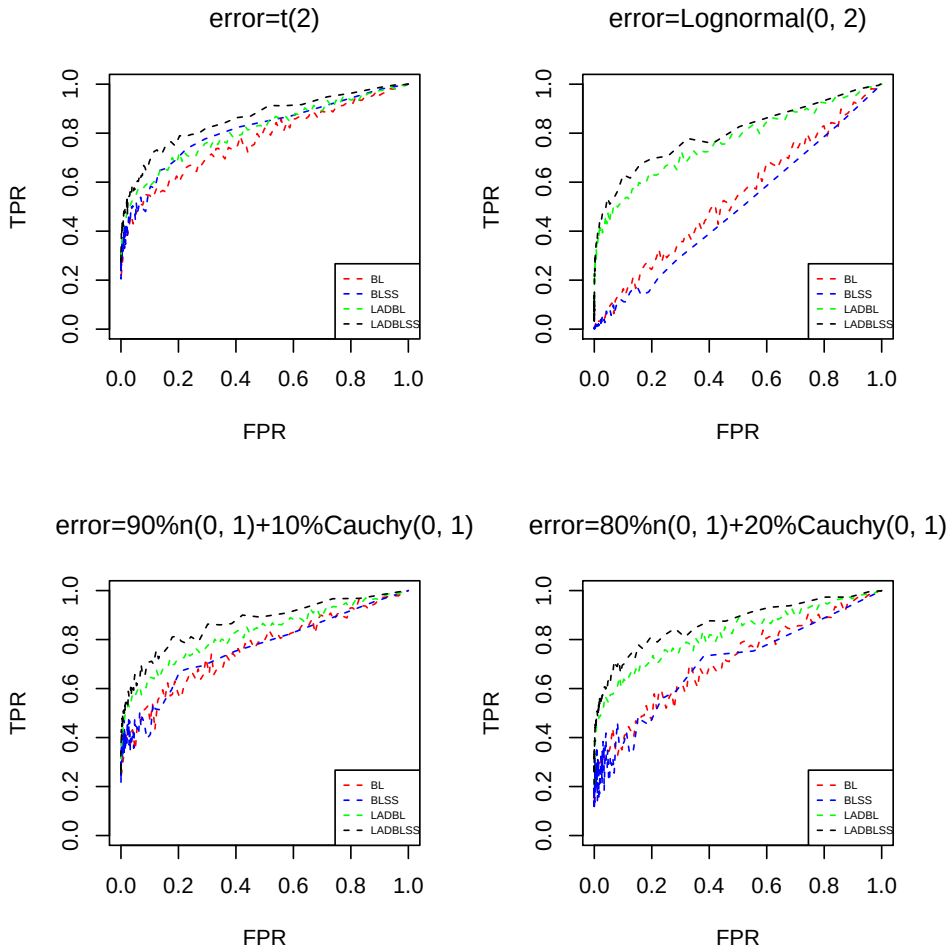

Figure: ROC curves of case2

Identification results

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

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

Identification results

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

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

Identification results

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

		Main	Interaction	Total
Error 1	BL	7.83(0.46)	3.80(1.09)	11.63(1.10)
N(0,1)	BLSS	7.90(0.31)	5.07(1.41)	12.97(1.35)
	LADBL	7.90(0.40)	3.60(1.49)	11.50(1.63)
	LADBLSS	7.33(0.52)	4.03(1.59)	11.76(1.48)
Error 2	BL	6.90(1.88)	1.87(1.48)	8.77(2.75)
$t(2)$	BLSS	6.97(1.27)	3.00(1.74)	9.97(2.35)
	LADBL	7.70(0.70)	2.93(1.46)	10.63(1.52)
	LADBLSS	7.47(0.82)	3.9(1.29)	11.37(1.59)
Error 3	BL	0.53(0.89)	0.53(0.73)	1.07(1.08)
Lognormal(0,2)	BLSS	0.63(0.81)	0.77(1.00)	1.40(1.19)
	LADBL	7.10(1.03)	1.60(1.25)	8.70(1.66)
	LADBLSS	6.73(1.20)	2.97(1.13)	9.70(1.95)
Error 4	BL	6.73(2.48)	1.87(1.36)	8.60(3.45)
90%N(0,1)	BLSS	6.10(2.86)	3.00(2.12)	9.10(4.51)
+10%Cauchy(0,1)	LADBL	7.97(0.18)	3.17(1.78)	11.13(1.74)
	LADBLSS	7.77(0.57)	4.53(1.69)	12.30(1.84)
Error 5	BL	6.13(2.33)	2.27(1.34)	8.40(2.91)
80%N(0,1)	BLSS	4.00(3.18)	1.43(1.50)	5.43(4.46)
+20%Cauchy(0,1)	LADBL	7.8(0.41)	2.73(1.44)	10.53(1.50)
	LADBLSS	7.80(0.48)	4.37(1.38)	12.17(1.51)

Summary

- LADBLSS outperforms all the other three approaches;
- Robust approaches performs well when heavy–tailed errors exist;
- Similar patterns have been observed across differents scenarios.

Applications to SKCM Data

- Data from TCGA-SKCM(The Cancer Genome Atlas Program-skin cutaneous melanoma)
- Top 10,000 genes with the strongest association with the response variables are chosen for downstream analysis (n=298).
- Response variable: log-transformed Breslow's thickness.
- Four environment factors: Age, AJCC pathologic tumor stage, gender and Clark level.
- To identify genes that have significant genetic main effect or $G \times E$ interaction effects on the Breslow's thickness.

Applications to SKCM Data

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

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

Results

Table: Analysis of the SKCM using approach LADBLSS.

Gene	Main Effects	Clark level	Stage	Age	Gender
EBF2	-0.0061				
KCTD16	-0.0345				
GH2	-0.635		-0.8708		
AVP		0.1131			
LOC105379569		0.1695			-0.3004
MTE		0.8074		-0.0839	-0.6608
UBQLN3			-0.3885		1.341
TDRG1				0.0617	
ACTL9		0.7542			
OR2T12		0.4096			
OR51A7		0.9043			
LINC00483		0.8008			
OR8I2	-0.6614	0.2902			
KCNK10				0.0096	
CRYAA		0.017			
H2A		0.0097			
OR4K14		1.6838	-0.4658	-0.4519	0.7227
LOC100133184		0.0084			
LOC100128288		0.0056			
HND		0.1046			
ODF1		0.6037			
DKKL1		0.0054			

Summary

- Propose a robust Bayesian Lasso with spike-and-slab priors approach to detect main and gene-environment interactions in a marginal model.
- Extensive simulation studies under different settings indicate the advantage of the method over the alternatives.
- The findings in case study are important for generating biological hypothesis for future lab validation.

Thank you for your attention!