IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 March 2024, accepted 28 March 2024, date of publication 2 April 2024, date of current version 15 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384309

RESEARCH ARTICLE

Sparse Variable Selection on High Dimensional Heterogeneous Data With Tree Structured Responses

HUI LIU[1,*], XIANG LIU[2,*], JING DIAO[3], WENTING YE[4], XUELING LIU[5], AND DEHUI WEI[6], (Graduate Student Member, IEEE)

[1]School of Public Administration, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
[2]School of Computing, National University of Singapore, Singapore 119077
[3]Department of Preventive Dentistry, Peking University School and Hospital of Stomatology, Beijing 100081, China
[4]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[5]Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597
[6]State Key Laboratory of Networking and Switching Technology, BUPT, Beijing 100876, China

Corresponding author: Hui Liu (huiliu.jxnu@gmail.com)

[*]Hui Liu and Xiang Liu contributed equally to this work.

ABSTRACT We consider the problem of sparse variable selection on high dimension heterogeneous data sets, which has been taking on renewed interest recently due to the growth of biological and medical data sets with complex, non-i.i.d. structures and huge quantities of response variables. The heterogeneity is likely to confound the association between explanatory variables and responses, resulting in enormous false discoveries when Lasso or its variants are naïvely applied. Therefore, developing effective confounder correction methods is a growing heat point among researchers. However, ordinarily employing recent confounder correction methods will result in undesirable performance due to the ignorance of the convoluted interdependency among response variables. To fully improve current variable selection methods, we introduce a model, the tree-guided sparse linear mixed model, that can utilize the dependency information from multiple responses to explore how specifically clusters are and select the active variables from heterogeneous data. Through extensive experiments on synthetic and real data sets, we show that our proposed model outperforms the existing methods and achieves the highest ROC area.

INDEX TERMS Confounding factors, genome-wide association study, mixed model, variable selection.

I. INTRODUCTION

Variable selection is one of the central tasks in statistics and has been studied for decades [1], [2]. Modern machine learning problems, especially biological or medical applications often seek solutions in the existing statistical approaches. Lasso [3] is an example of those widely adopted methods in a variety of areas for sparse variable selection tasks. However, the increasing volume of data sets often requires the data to be collected from multiple batches and then integrated together. This procedure is particularly harmful to the biological [4] and medical [5], [6] data sets, which are sensitive to the data sources, like populations, hospitals or even experimental devices. This sensitivity results in the heterogeneity, therefore, breaks one of the most fundamental assumptions (i.i.d. assumption) that most statistical machine learning methods make. More importantly, due to the expensiveness of biological and medical data, different batches of data are gathered for different purposes from distinctly different sources, such as samples of the control group are mostly collected from volunteers from several different undeveloped regions. Consequently, the heterogeneity often induces confounding factors between explanatory variables and response variables, resulting in numerous false positive selected variables when classical variable selection techniques are applied [7].

To deepen understanding of the challenge that heterogeneity is introduced to biological or medical data sets and define the problem, consider that we have data samples in the format of (X, Z, Y), where X stands for the explanatory variables, Y

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

stands for the responses, Z stands for the indicator of the data source. The dependency between X and Y is the premise of any variable selection tasks [8], and the dependency between X and Z is induced through heterogeneity [9], [10]. The data collection procedure we mentioned brings the dependency between Z and Y. In the real world, this problem may be even intractable, for the origin of different samples is lost either through data compression or experimental necessity in most cases. Nowadays genetic association studies are rarely aware of the origin of the samples listed. Z becomes the confounding factor between X and Y [11], [12], [13], [14]. One challenge of the heterogeneous data variable selection problem is to mitigate the confounding effects brought by Z.

Aside from challenges above, many of the real world biological and medical data sets are collected along with multiple response variables. These responses are often more closely related and could share common relevant covariates than others and then form the tree or other kinds of structures [15], [16], [17], [18]. For instances, in genetic association analysis, which aims to select the single-nucleotide polymorphism (explanatory variables) that could affect the phenotype (response variables), the genes in the same pathway pretend to share the common set of relevant explanatory variables than other genes.

Thus, to improve the performance of the variable selection, incorporating the complex correlation structure in the responses is under our consideration. In this paper, we extend the recent solutions of sparse linear mixed model [8], [9] that can correct confounding factors and perform variable selection simultaneously further to account the relatedness between different responses. We propose the **t**ree-**g**uided **s**parse **l**inear **m**ixed **m**odel, namely TgSLMM, to correct the confounder and incorporate the relatedness among response variables simultaneously. With TgSLMM, we are capable to improve the performance of the variable selection when considering the statistical criterion, incorporating the complex tree-based correlation structure in the traits under our consideration. Eventually, we examine our model through plenty of repeated experiments and show that our method is superior to other existing approaches and able to discover the real genome association in the real data set.

II. RELATED WORK

Recent years have witnessed the great advances in the variable selection area. The most classical approach is ℓ_1-norm regularization (i.e. *Lasso* regression [3]). Further, studies have extended the model capability by introducing various regularizers [15]. Examples including the Smoothly Clipped Absolute Deviation (SCAD) [19], the Local Linear Approximation (LLA) [20], the Minimax Concave Penalty (MCP) [21], and the Precision Lasso [22] have been introduced since then, which all overcome a variety of limitations of Lasso [19]. Some other variable selection methods like [23] ignore underlying multidimensional structure, leading to severe small dataset problems. Reference [24] imposes a rank constraint into ℓ_1 regularization to factor matrices and promotes sparsity in variable selection, which hurts the interpretability. The liability-threshold mixed linear model overcomes the limitation of Linear Mixed Model (LMM) in case-control ascertainment [25]. Reference [26] proposed a unsupervised variable selection method. But both of them cannot apply to high dimensional data with heterogeneity.

Besides these, in the non-i.i,d setting, confounders could raise a challenge in variable selection when the data set is originated from different sources. Corresponding solutions have been studied for decades. Principal components analysis (PCA) [27], [28] and linear mixed model [29], [30] are two popular and efficient approaches to alleviate the confounding effect. The latter provides a more fine-grained way to model the population structure and won its prominence in the animal breeding literature, where it was used to reveal the underlying kinship and family structure [11], [31]. Many extensions have been developed, however, these measures such as LMM-Select [32] LMM-BOLT [33] and Liability-threshold mixed linear model (LTMLM) [33] along with other algorithms [34], [35], [36] only rely on univariate testing to select the variable once uncovering the confounding factor. Attempts have been made to propose multi-variable testing model [8], [9], [37], [38] these days, but their performances fall short while tackling with the challenge that takes the relatedness between responses into account. References [39], [40], [41], [42], [43], [44], and [45] are proposed to identify significant associations, which is to be contrasted with the related problem of estimating heritability. However, they also lack accounting for the relatedness between different traits [46]. Reference [47] helps improve association methods for kinship estimation, but it could not construct the convoluted phenotypic architecture in a dataset originated from different populations in the real world like [16]. The challenges show the desire to have a method, which requires no prior knowledge of the individual relationship and is capable of uncover the structured pattern in a way that is properly calibrated to the degrees of traits' relatedness.

III. TREE-GUIDED SPARSE LINEAR MIXED MODEL

Throughout this paper, X denotes the $n \times p$ matrix for explanatory variables for individuals, Y denotes the $n \times k$ matrix for response variables, and β denotes the $p \times k$ matrix for effect sizes. We use subscripts to denote rows and superscripts to denote columns, for example, β_k and β^k are the k-th column and k-th row of β respectively.

In this section, we begin by examining the sparse linear mixed model. Next, we demonstrate how we leverage the technique to uncover relationships between traits. Finally, we transform this approach into a regression problem and employ efficient methods to address associated challenges.

A. SPARSE LINEAR MIXED MODEL

The linear mixed model (LMM) is an extension of the standard linear regression model that explicitly describes the relationship between response variables and explanatory variables incorporating an extra random term to account for confounding factors. To introduce the sparse linear mixed model, we briefly revisit the classical linear mixed model as Equation 1:

$$Y = X\beta + Zu + \epsilon \tag{1}$$

where Z is an $n \times t$ matrix for the random effect. u is the confounding influences with implicitly identity correlation information, ϵ denotes observed noise and they both follow the independent Gaussian distribution with the zero means. Intuitively, Zu models the covariance between the observations y_i. Assuming that $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, $u \sim \mathcal{N}(0, \sigma_g^2 I)$, $K = ZZ^T$ and represents the covariance between the responses and σ_g represents the magnitude of confounder factors, we can rewrite the formula as Equation 2 to simplify mathematical derivation:

$$y_k \sim \mathcal{N}(X\beta_k, \sigma_g^2 K + \sigma_\epsilon^2 I) \tag{2}$$

Assuming the priori distribution of β could be expressed as $e^{-\Phi(\beta)}$, we can define log likelihood function as Equation 3:

$$\ell(\sigma_g^2, \sigma_\epsilon^2, \beta) = e^{-\Phi(\beta)} \cdot \prod_{k=1}^{K} \mathcal{N}(y_k | X\beta_k, \sigma_g^2 K + \sigma_\epsilon^2 I) \tag{3}$$

Based on the sparsity of β, it's reasonable to assume that β follows Laplace shrinkage prior. Such assumptions lead to the sparse linear mixed model. However, sparse LMM fails to consider the relatedness among response variables. The defect drives us to the tree-guided sparse linear mixed model.

B. TREE-GUIDED SPARSE LINEAR MIXED MODEL

To incorporate the relatedness among responses simultaneously, we use Tree-Lasso as Equation 4.

$$\Phi(\beta) = \lambda \sum_j \sum_{v \in V} w_v ||\beta_j^{G_v}||_2 = \lambda \sum_j W_j(v_{root}) \tag{4}$$

where λ is a tuning parameter that controls the amount of sparsity in the solution and $\beta_j^{G_v}$ is a vector of regression coefficients $\{\beta_j^k | k \in G_v\}$. The overlaps of groups of Tree-Lasso and the number of the trees is determined by the hierarchical clustering tree. Each node $v \in V$ of the j-th tree is associated with the group G_v whose members are the response variables at the nodes of the same subtree. Each group of subtree regression coefficients $\beta_j^{G_v}$ is weighted with w_v, which is defined as the Equation 5. In general, h_v in the Equation 5 represents the weight for selecting relevant covariates separately for the responses associated with each child of node v, whereas the $1 - h_v$ represents the weight for selecting relevant covariates jointly for the responses for all of the children of node v, and the value of h_v ranges from 0 to 1. Assuming K is the number of response variables and $|V|$ is equivalent to the number of nodes in one tree, since a tree associated with K responses has $2K - 1$ nodes, $|V|$ appears in the tree-lasso penalty is upper-bounded by $2K$.

$$w_v = \begin{cases} (1 - h_v) \prod\limits_{m \in Ancestors(v)} h_m & \text{if v is an internal node,} \\ \prod\limits_{m \in Ancestors(v)} h_m & \text{if v is a leaf node.} \end{cases} \tag{5}$$

To simplify the computation process of Tree-Lasso, we can calculate the separate penalty from the root of each tree iteratively as Equation 6:

$$W_j(v) = \begin{cases} (1 - h_v) \sum\limits_{c \in Children(v)} |W_j(c)| + h_v ||\beta_j^{G_v}||_2, \\ \text{if v is an internal node.} \\ \sum\limits_{m \in G_v} |\beta_j^m|, \\ \text{if v is a leaf node.} \end{cases} \tag{6}$$

C. PARAMETER LEARNING

Overall, optimizing Equation 3 with hyper-parameter $\{\Theta = \sigma_g^2, \sigma_\epsilon^2, \lambda, w_v\}$ is a non-convex optimization problem aside with weights β. Hence, we could apply the null-model fitting method first to correct the confounding factors and then solve Tree-Lasso regression problem using smoothing proximal gradient method [48].

1) NULL-MODEL

Due to sparsity of β, null-model fitting method by first optimizing $\sigma_g^2, \sigma_\epsilon^2$ while ignoring individual explanatory variables, can yield near-identical result as an exact method [30]. By using the computational trick [34] that introduces the ratio of the random effect and the noise variance, $\delta = \sigma_\epsilon^2 / \sigma_g^2$, we could transform the equation as Equation 7:

$$\ell_{null}(\sigma_g, \delta) = e^{-\Phi(\beta)} \cdot \prod_{k=1}^{K} \mathcal{N}(y_k | X\beta_k, \sigma_g^2(K + \delta I)) \tag{7}$$

The genetic effects are treated as fixed effects, whereas the confounding influences are modeled as random effects. We carry out a log likelihood optimization with regard to δ and then σ_g in closed form.

2) REDUCTION TO TREE-LASSO REGRESSION PROBLEM

In general, we first compute the spectral decomposition of $K = U\text{diag}(d)U^T$, where U for eigenvector matrix and diag(d) for eigenvalue matrix. Having the yielded δ, we use the U to reweight the data such that the covariance matrix becomes isotropic. Assume $\tilde{Y}$ and $\tilde{X}$ are the resulting rescaled data, which can be calculated by the following equation:

$$\tilde{X} = (\text{diag}(d) + \delta I)^{-\frac{1}{2}} U^T X$$
$$\tilde{Y} = (\text{diag}(d) + \delta I)^{-\frac{1}{2}} U^T Y$$

Using this transformation, the equation eventually ends up with a standard Tree-Lasso regression problem since it

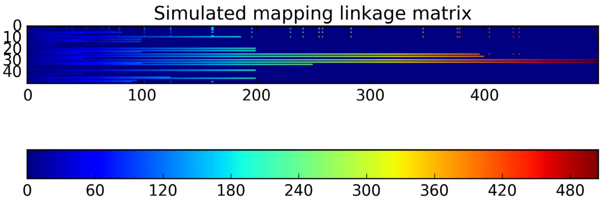


FIGURE 1. The simulated ground-truth β vector. For the illustration purpose, we choose the experimental setting of $n = 250$, $p = 500$ and $k = 50$.

is free of population structure and has been alleviated the confounding factor. In the following step, we can obtain the $\widehat{\beta}^{tree}$ as Equation 8:

$$\widehat{\beta}^{tree} = \min_{\beta} \frac{1}{2}||\tilde{Y} - \tilde{X}\beta||_F^2 + \Phi(\beta) \tag{8}$$

where $|| \cdot ||_F$ denotes the matrix Frobenius norm, and Φ is determined by the Equation 4, then we can easily employ the smoothing proximal gradient descent method.

IV. SYNTHETIC EXPERIMENTS

In this section, we evaluate the yielded results of the TgSLMM versus Tree-Lasso, LMM-Lasso and some techniques mentioned above, which is shown in the receiver operating characteristic (ROC) curves.[1]

A. DATA GENERATION

First, we simulate a sparse tree-structured vector as β. An illustrated example is shown in Figure 1. To construct β, the generation rules are listed below:

- The righter columns have fewer non-zero elements.
- The elements from righter columns have bigger value.
- Some non-zero elements are shattered discretely in β to increase the complexity and mimic the real situation.

Then we generate centroids of m different distributions. With c_j as the centroid of j-th distribution, we generate explanatory variable data from a multivariate Gaussian distribution as follows:

$$x_i \sim \mathcal{N}(c_j, \sigma_e^2 I) \tag{9}$$

where x_i denotes the i-th data or information bore by one individual and originates from j-th distribution chosen from m different distributions c. Then we generate an immediate response vector r from X matrix with $\epsilon \sim (0, \sigma_\epsilon^2)$:

$$r = X\beta + \epsilon \tag{10}$$

To get the final response matrix Y, we introduce a covariance matrix K to simulate correlation between different responses:

$$Y \sim \mathcal{N}(r, \sigma_y^2 K) \tag{11}$$

TABLE 1. Default experimental setting in the simulated experiments.

Parameter	Default	Description
n	1000	the number of data samples
p	5000	the number of explanatory variables
k	50	the number of response variables
m	10	the number of distributions that data originates from
d	0.05	the percentage of active variables
σ_e^2	0.001	the magnitude of covariance of explanatory variables
σ_y^2	1	the magnitude of covariance of response variables
σ_ϵ^2	0.05	the magnitude of noise

where σ_y^2 is to control the magnitude of the variance. Assuming C is the matrix formed by stacking the centroids c_j, we choose $K = CC^T$ to simulate the correlation between observations.

Using the data generation method described earlier, our synthetic dataset can effectively mimic real-world heterogeneous datasets, capturing the desired trait relatedness.

B. EXPERIMENT RESULTS

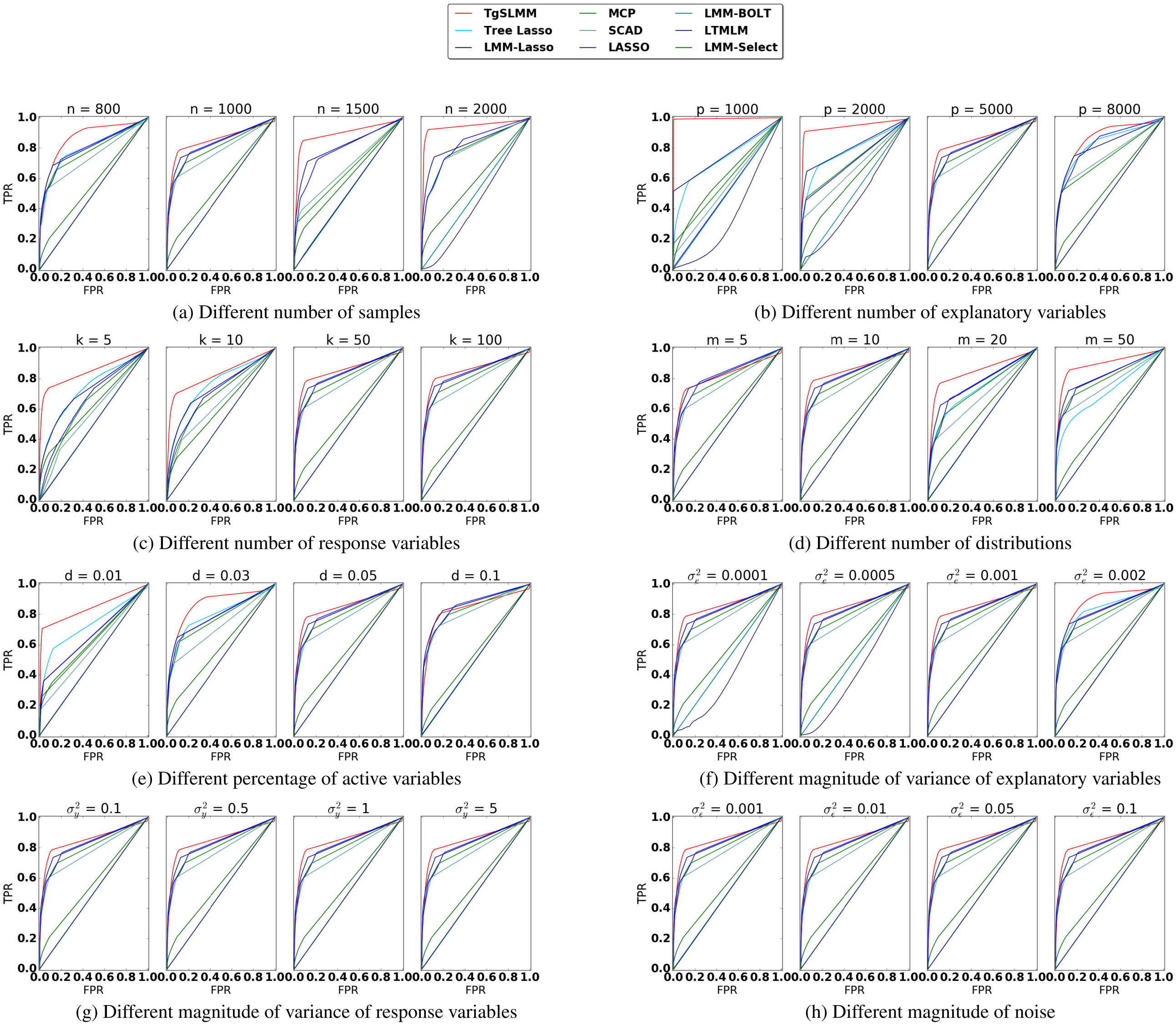
We assessed the ability of TgSLMM in our synthetic data sets. The experimental setting is listed in Table 1.

To evaluate the performance of the proposed model in identifying active variables in different data sets, Tree-Lasso, LMM-Lasso, MCP, SCAD, Lasso,[2] BOLT-LMM, LTMLM and LMM-Select are also tested. The baselines we choose in this paper are all highly cited and have been proven effective by many scholars. In general, our method exceeds all the other methods. The results are shown in Figure 2 considering the golden criterion ROC curves.

In Figure 2(a) as n increases, and in Figure 2(b) as p decreases, the ratio of $\frac{p}{n}$ gets smaller and the performance gets better as expected. Compared to Tree-Lasso along with other methods, our method is more robust with big data sets, which suits the real-world situation. As we increase the number of response variables in Figure 2(c), increase the number of distributions in Figure 2(d), or decrease the proportion of active variables in β as Figure 2(e), the problem becomes more challenging. Figure 2(f) and Figure 2(g) show that our method is more flexible to different magnitudes of covariance of explanatory variables and response variables. In Figure 2(e), we notice that when the proportion of active variables in β is large, the performance of TgSLMM and LMM-Lasso is similar. However, it contradicts the background of our research that the active variables should be sparse among data. Through our experiments, it is hard

[1] The problem can be regarded as classification problem–identifying the active response variables from all genes. For each threshold, we select the response variables whose absolute effect sizes are greater than the threshold. If the selected explanatory variable has value above the threshold in ground truth effect size, it will be the true positive.

[2] We modify Lasso, MCP, SCAD to support multidimensional data processing, the performance they yield has no observed difference. Our contributions and the details can be viewed in our codes.



(a) Different number of samples

(b) Different number of explanatory variables

(c) Different number of response variables

(d) Different number of distributions

(e) Different percentage of active variables

(f) Different magnitude of variance of explanatory variables

(g) Different magnitude of variance of response variables

(h) Different magnitude of noise

FIGURE 2. ROC curves for experiments with different parameters. We show the full image of ROC curves to compare our method with previous methods. For each configuration, the reported curve is drawn over five random seeds.

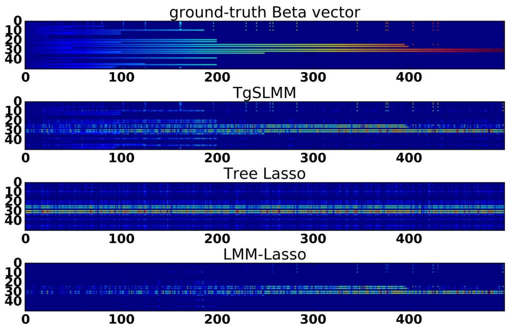


FIGURE 3. The yielded β vectors.

for Tree-Lasso to identify the active variables on high dimensional heterogeneous data.

TgSLMM also performs best in most cases in the figure of Precision-Recall curves. These figures are shown in Appendix.

C. ANALYSIS OF YIELDED β AND Y

We use the same experimental setting[3] as in Figure 1 to perform the ablation studies. The results are shown in Figure 3 and 4.

Figure 3 shows that TgSLMM recovers both the values and structure of ground truth effect size, revealing the supreme ability of TgSLMM in variable selection. LMM-Lasso has trouble finding enough useful information. Trapped into the confounding factors, the Tree-Lasso discovers too

[3] Other parameters are as follow: m is 10; d is 0.05; σ_e^2 is 0.001; σ_y^2 is 1; σ_ϵ^2 is 0.05.

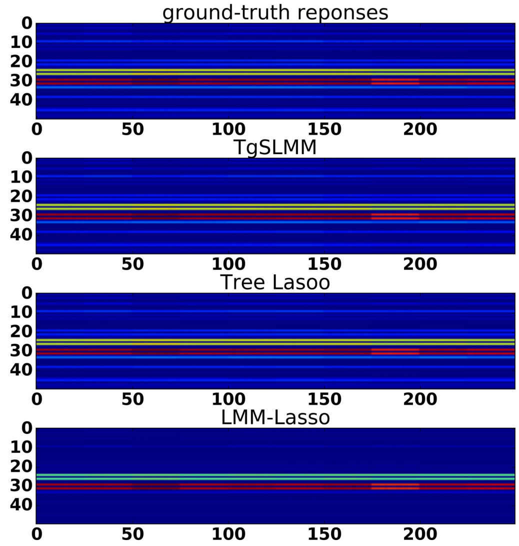


FIGURE 4. The simulated responses matrices and the predicted responses results by different models.

many false positives. Tree-Lasso also falls short when the data set becomes complicated in the Figure 2. Based on Figure 4, both prediction performance of TgSLMM and Tree-Lasso are convincing, LMM-Lasso fails as reported before. Unsurprisingly, the proposed TgSLMM also behaves the best in estimating β with respect to mean-squared error through almost all the experimental settings. The other approaches cannot discover any meaningful information.

By using the proposed method, we are able to detect weak signals and reveal clear groupings in the patterns of associations between explanatory variables and responses and apply our method to many applications, such as variable selection, effect sizes estimation, and response prediction.

V. REAL GENOME DATA EXPERIMENTS

Having shown the capacity of TgSLMM in recovering explanatory variables of synthetic data sets, we now demonstrate how TgSLMM can be used in real-world genome data and discover meaningful information. To evaluate the method, we focus on some practical data sets, Arabidopsis thaliana, Heterogeneous Stock Mice and Human Alzheimer Disease. Since Arabidopsis thaliana and Heterogeneous Stock Mice have been studied for over a decade, the scientific community has reached a general consensus regarding these species [49]. With such authentic golden standard, we could plot the ROC curve and assess the model's performance using the area under it. However, since Alzheimer's disease is a very active area of research with no ground truth available, we list the genetic variables identified by our proposed model and verify the top genetic variables by searching the relevant literature.

A. DATA SETS

1) ARABIDOPSIS THALIANA

The Arabidopsis thaliana data set we obtained is a collection of around 200 plants, each with around 215,000 genetic variables [50]. We study the association between these genetic variables and a set of observed responses. These plants were gathered from 27 different countries in Europe and Asia, so that geographic origin served as a potential confounding factor. For example, different sunlight conditions in different regions may affect the observed responses of these plants. We test the genetic associations between genetic variables with 44 different responses such as days to germination, days to flowering, *etc*.

2) HETEROGENEOUS STOCK MICE

The heterogeneous stock mice data set contains measurements from around 1700 mice, with 10,000 genetic variables [51]. These mice were raised in cages by four generations over a two-year period. In total, the mice came from 85 distinct families. The obvious confounding variable is genetic inheritance due to family relationships. We study the association between the genetic variables and a set of 27 response variables that could possibly be affected by inheritance. These 27 response variables fall into six different categories, relating to the glucose level, insulin level, immunity, EPM, FN and OFT respectively.

3) HUMAN ALZHEIMER DISEASE

We use the late-onset Alzheimer's Disease data provided by Harvard Brain Tissue Resource Center and Merck Research Laboratories [52]. It consists of measurements from 540 patients with 500,000 genetic variables. We test the association between these genetic variables and 28 responses corresponding to a patient's disease status of Alzheimer's disease.

4) PREPROCESSING OF REAL GENOMIC DATA

Each element of the explanatory variables X takes values from $\{0,1\}$ according to the number of minor alleles frequency (MAF) at the given locus in each individual. We also standardized the traits data, according to the analysis and statistics law. In the experiments, we found that the standardizing process is very crucial to the performance of the model.

B. ARABIDOPSIS THALIANA

Since we have access to a validated gold standard of the Arabidopsis thaliana data set, we compare the alternative algorithms in terms of their ability in recovering explanatory variables with a true association. Figure 5 illustrates the area under the ROC curve for each response variable for Arabidopsis thaliana. By analyzing the results, we conclude that TgSLMM equals or exceeds the other methods for all of responses. TgSLMM allows for dissecting individual explanatory variable effects from global genetic effects driven by population structure.

Further, we simply apply linear regression and cross-validation to evaluate the proposed model's ability of response prediction versus all the algorithms. Using the

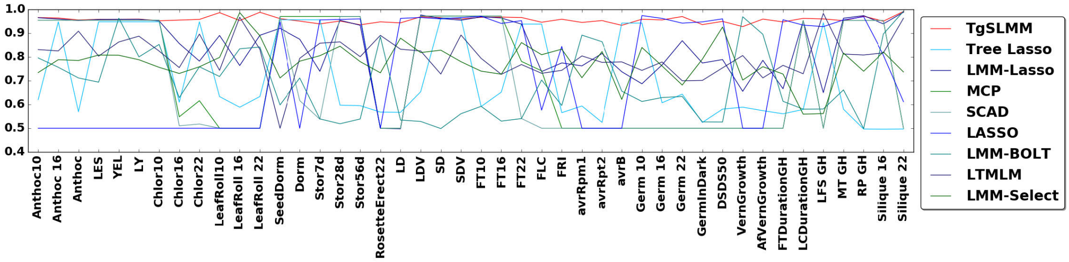


FIGURE 5. Area under ROC curve for the 44 traits of Arabidopsis thaliana.

TABLE 2. Discovered genetic variables with TgSLMM.

Rank	SNP	Rank	SNP	Rank	SNP
1	rs30882	34	rs10775247	67	rs3129317
2	rs10027921	35	rs13129773	68	rs7588938
3	rs12641981	36	rs9598119	69	rs551499
4	rs12506805	37	rs3767444	70	rs11776648
5	rs684240	38	rs10904362	71	rs635154
6	rs16844380	39	rs7516808	72	rs10965041
7	rs6431428	40	rs7039420	73	rs2301230
8	rs12509328	41	rs11673516	74	rs164415
9	rs7783626	42	rs6552578	75	rs11203999
10	rs11848278	43	rs7043499	76	rs7302430
11	rs10897029	44	rs11642659	77	rs31331
12	rs464906	45	rs6571869	78	rs1421203
13	rs874404	46	rs12648715	79	rs11953877
14	rs4421632	47	rs1475668	80	rs11607862
15	rs6086773	48	rs10247315	81	rs6762993
16	rs4882754	49	rs467089	82	rs7969169
17	rs2272445	50	rs10512516	83	rs12523589
18	rs12410705	51	rs17077288	84	rs6544170
19	rs4578488	52	rs6852162v	85	rs8107465
20	rs3887171	53	rs2492303	86	rs509936
21	rs2298955	54	rs12664420	87	rs659505
22	rs1998933	55	rs993312	88	rs2415449
23	rs17467420	56	rs4890939	89	rs17173637
24	rs7629705	57	rs1979687	90	rs2519126
25	rs27162	58	rs4702249	91	rs11605879
26	rs4740820	59	rs10233816	92	rs812462
27	rs1495805	60	rs6118709	93	rs13250449
28	rs7916633	61	rs2798639	94	rs1080310
29	rs13221797	62	rs12345602	95	rs4570478
30	rs429536	63	rs4766333	96	rs4787760
31	rs7668750	64	rs3814391	97	rs2682585
32	rs1463118	65	rs1008411	98	rs2854439
33	rs1163825	66	rs12815078	99	rs9301747

explanatory variables our proposed method selects, 61.4% of prediction for Arabidopsis thaliana is better than using origin data set, 56.8% is better than using the data after employing Tree-Lasso, 79.5% is better than applying LMM-Lasso, 84.1% is better than MCP and SCAD, 66.0% is better than Lasso, 91.0% is better than LMM-BOLT, 56.7% is better than LTMLM. Our method only works worse than LMM-Select while considering prediction.

C. HETEROGENEOUS STOCK MICE

For Heterogeneous Stock Mice data set, ground truth is also available so that we could evaluate the methods based on the area under their ROC Curve as Figure 6. TgSLMM behaves as the best one on 22.2% of the traits and achieves the highest ROC area for the whole data set as 0.627. The second best model is MCP with the area of 0.604. The areas under ROC of Tree-Lasso, Lasso and SCAD are 0.582, 0.591 and 0.590 respectively. The areas of the remaining models are all around 0.5, showing little ability to process such complex data sets. On traits Glucose_75, Glucose_30, Glucose.DeadFromAnesthetic, Insulin.AUC, Insulin.Delta and FN.postWeight, our method TgSLMM behaves the best. The results are interesting: the left side of the figure mostly consists of traits regarding glucose and insulin in the mice, while the right side of the figure consists of traits related to immunity. This raises the inspiring question of whether or not immune levels in stock mice are largely independent of family origin.

D. HUMAN ALZHEIMER DISEASE

Finally, we proceed to the Human Alzheimer's Disease data set and report the top 99 genetic variables our model discovered in Table 2 to foster further research.

Due to space limitation, we only verify the top 10 reported genetic variables with prior research. The 1^{st} discovered genetic variable is corresponded to *apoB* gene, which can influence serum concentration [53] in Alzheimer's disease [54]. The 2^{nd} one is associated with *ARHGAP10* gene (also called *GRAF2*), which is an important paralogon of *ARHGAP26* that closely related to the Alzheimer's disease [55] and affects the developmentally regulated expression of the *GRAF* proteins that promote lipid droplet clustering and growth, and is enriched at lipid droplet junctions [54], [56]. The 3^{rd} SNP *GNPDA2* is discovered to show the environment and gene association with obesity. They

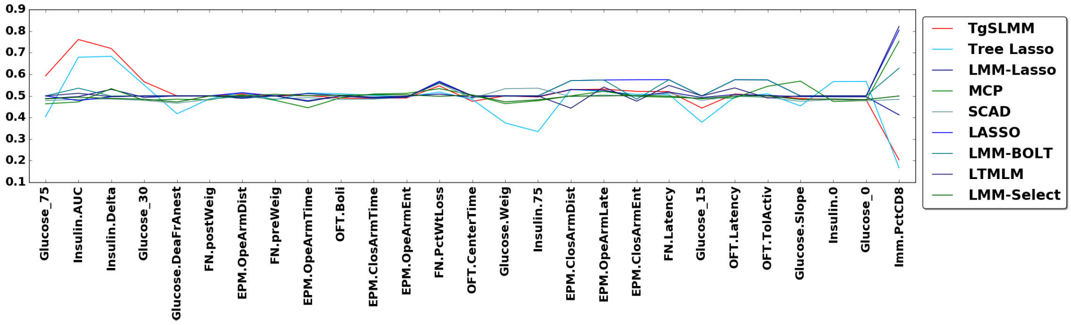
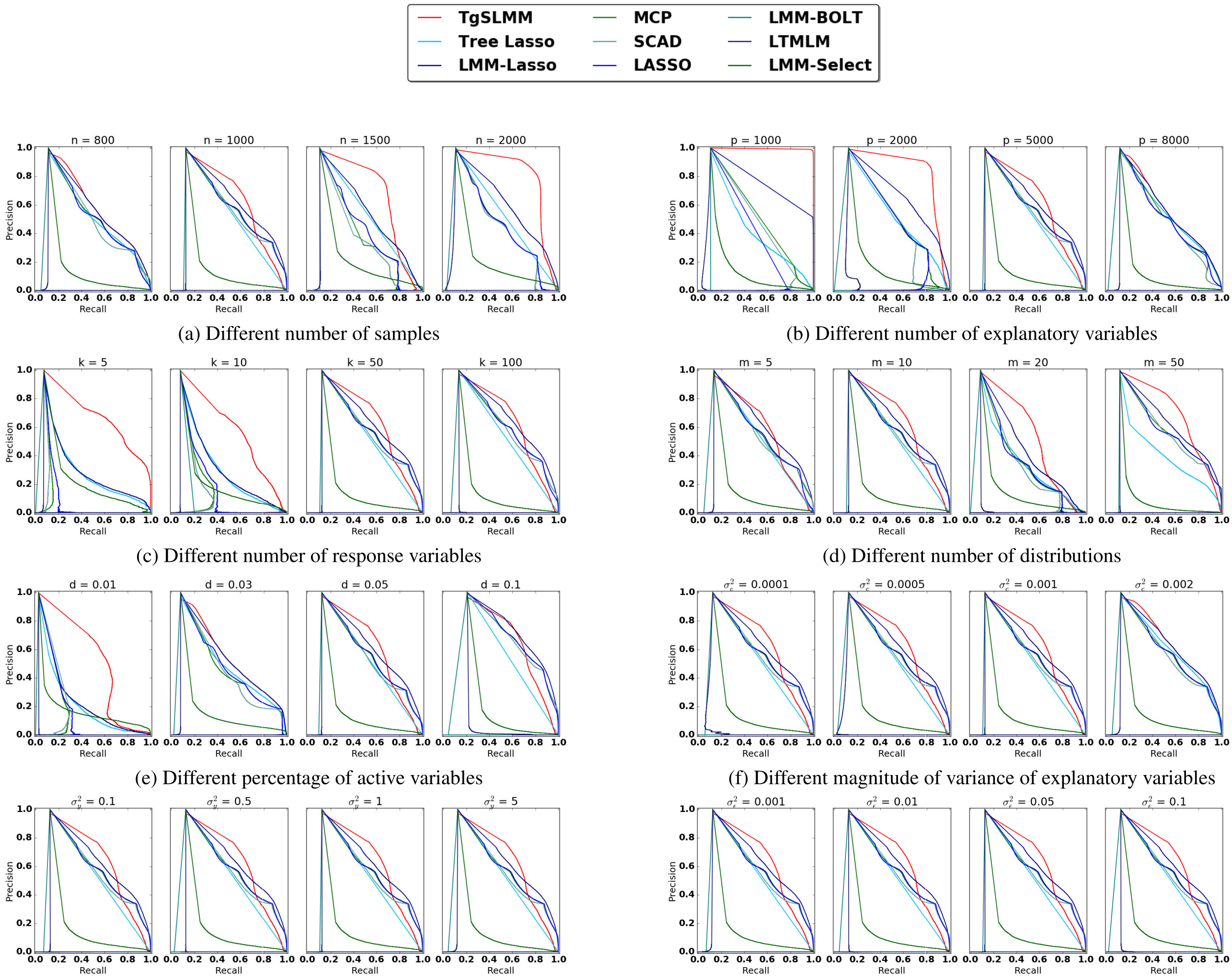


FIGURE 6. Area under ROC curve for the 27 traits of mice.



(a) Different number of samples

(b) Different number of explanatory variables

(c) Different number of response variables

(d) Different number of distributions

(e) Different percentage of active variables

(f) Different magnitude of variance of explanatory variables

(g) Different magnitude of variance of response variables

(h) Different magnitude of noise

FIGURE 7. Precision-Recall curves for experiments with different parameters.

have impact on neurodegenerative and neurodevelopmental diseases [57]. The 4^{th} SNP is expressed by the *SYNPO2*, which influences hypercholesterolemia or hypertension that has a identified a link between cognitive deficits [58]. The 6^{th} SNP known as the *LY75*, has close relation with the significantly differentially expression in the time-series paired analysis involving *APOE4* carriers and non-carriers, which could affect Alzheimer's disease [59]. The 7^{th} genetic variable is associated with *AGAP1*. *AGAP1* can regulate membrane trafficking, actin remodeling [60] and is reported

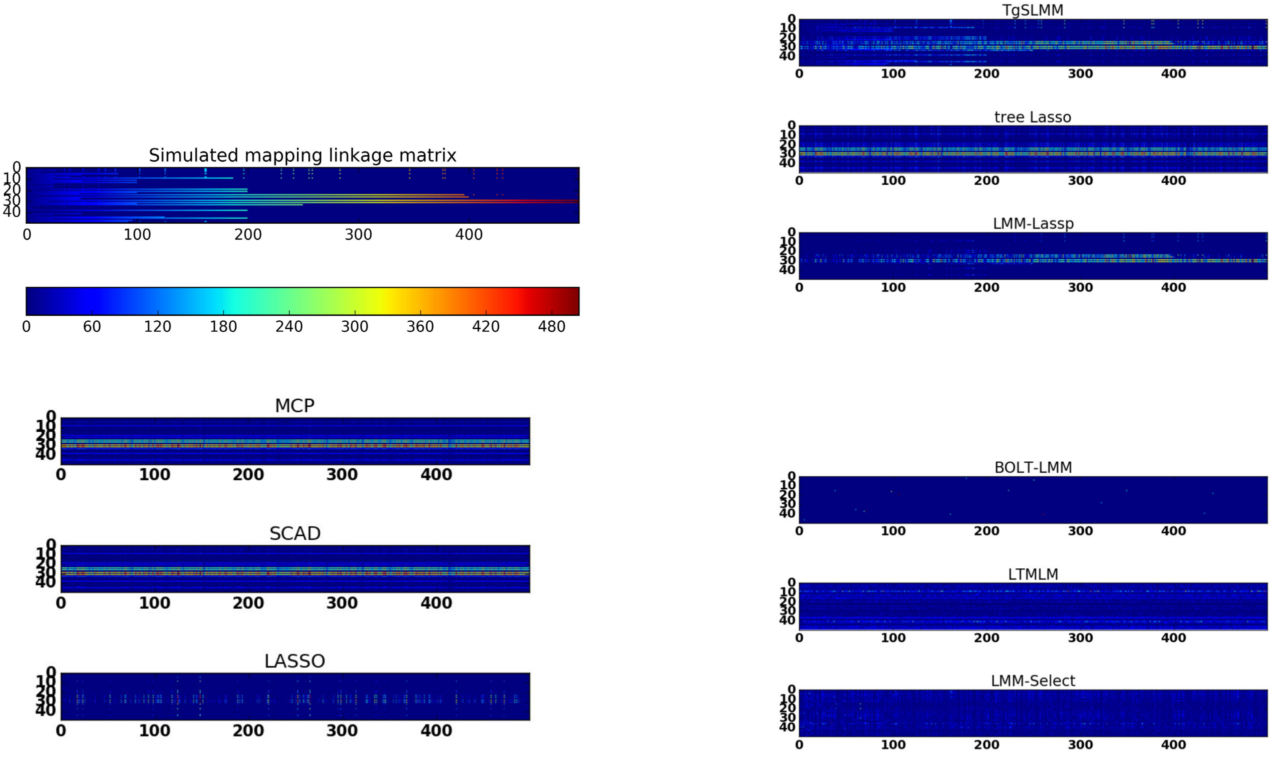


FIGURE 8. The yield β vector.

to be associated with Alzheimer's disease. The 8^{th} one is coded by gene *FAM114A1*. Biologists have found that *FAM114A1* is highly expressed in the developing neocortex [61]. Also, from "the amyloid hypothesis", beta-amyloid accumulation is mainly cause Alzheimer's disease [62]. The 9^{th} is corresponded with gene *CNTNAP2* and the direct downregulation of *CNTNAP2* by *STOX1A* is associated with Alzheimer's disease [63].

VI. DISCUSSIONS

A. COMPLEXITY

Since a tree associated with L responses can have at most $2L - 1$ nodes, it is computationally efficient and spatially economical to run TgSLMM. The complexity of TgSLMM is dependent on two parts. First, the decomposition of the random effect matrix K to rotate the explanatory variable and response data is cubic cost, which determines the complexity of the first step. If we reduce the covariance K to a low-rank representation calculated from a small subset of n_s explanatory variables. The runtime is reduced from $O(nk^2)$ to $O(n_s^2 k)$. Second, we employ a smoothing proximal gradient method that is originally developed for structured-sparsity-inducing penalties. By using the efficient method, the convergence rate of the algorithm is $O(\frac{1}{\epsilon})$, given the desired accuracy ϵ and the time complexity per iteration of the smoothing proximal gradient for the Tree-Lasso is $O(p^2k + p\sum v \in V|G_v|)$. Thus the overall complexity for our method is $O(n_s^2 k + \frac{1}{\epsilon} \times (p^2k + p\sum v \in V|G_v|))$.

B. RUNTIME

To evaluate its effectiveness and practicability, we have empirically measured the runtime on the Arabidopsis thaliana dataset mentioned in our paper. On a four-core computer (3GHz 12MB L2-Cache, 8GB Memory), TgSLMM required about 4 hours CPU time. In this paper, we show that our method is scalable to large genetic dataset.

VII. CONCLUSION

In this paper, we aim to solve the challenging task of sparse variable selection when the data are not i.i.d. This type of situation often occurs in genomics since different batches of medical data are collected from different sources for different purposes. Due to such confounding factors, naiv̈ely applying the traditional variable selection methods will result in a huge number of false discoveries. In addition to that, existing algorithms ignore the convoluted interdependency among responses, hence a joint analysis that can utilize such relatedness information in a heterogeneous data set is crucial. To address these problems, we propose the tree-guided sparse linear mixed model for sparse variable selection. Apart from extending the recent solutions of LMM that can correct confounding factors, we can perform variable selection simultaneously further to account the relatedness between different responses. By conducting extensive experiments, we compare our method with state-of-art methods and deeply analyze how confounding factors from the high dimensional heterogeneous data set influence the capability of the model

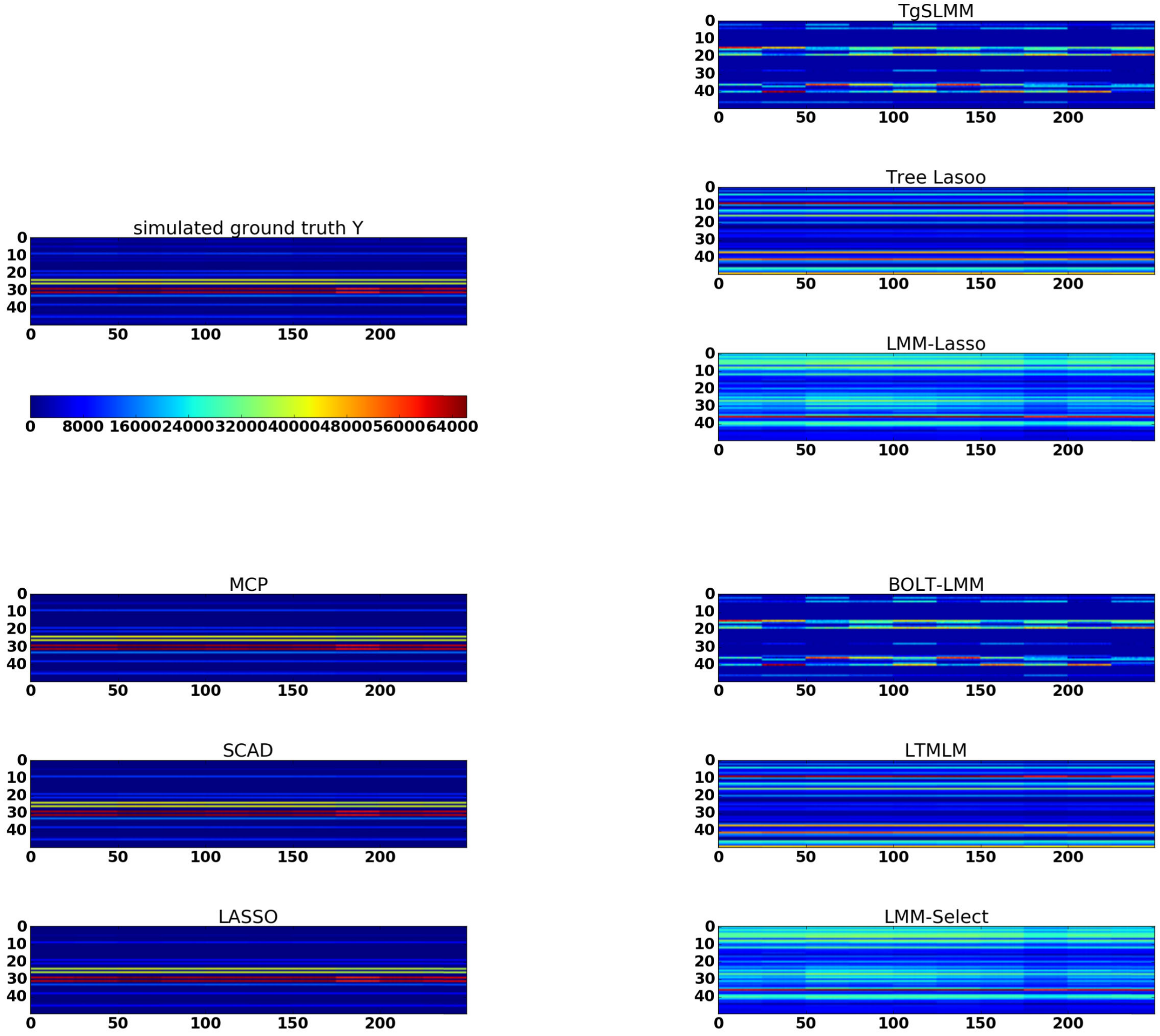


FIGURE 9. **The simulated responses matrix and the yield responses results.**

to identify active variables. We show that traditional methods easily fall into the trap of utilizing false information, whereas our proposed model outperforms other existing methods in both the synthetic data set and real genome data set. We make our source code available.[4]

APPENDIX A
SYNTHETIC EXPERIMENT RESULTS

A. THE PRECISION-RECALL CURVE OF SYNTHETIC EXPERIMENT

The Figure 7 shows the full images of Precision-Recall curves in synthetic experiments to compare our method with other existing methods by using the same parameters in our paper. For each configuration, the reported curve is drawn over five random seeds. And we can see that TgSLMM behaves almost always best.

B. ESTIMATION OF β

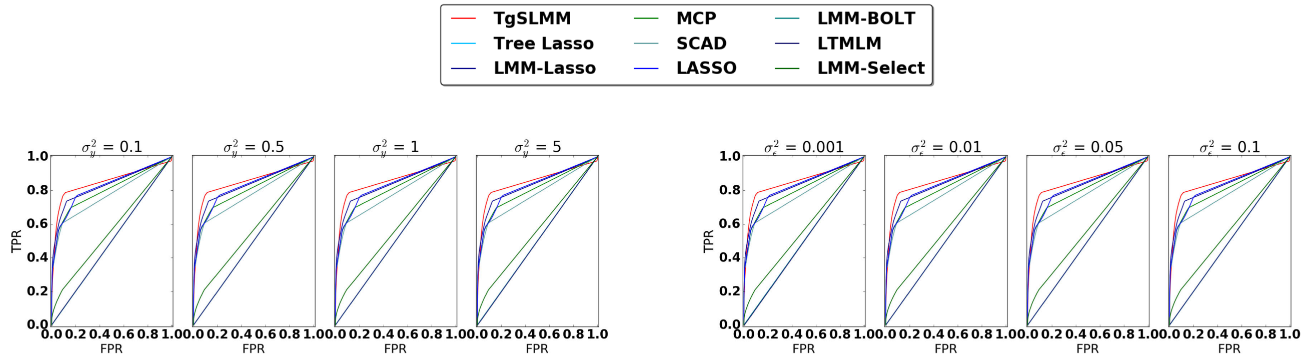
The Figure 8 shows the β vectors yielded by methods we used in our paper together with the ground truth β vector generated in the synthetic experiments. The figures show that TgSLMM yields the best result with the number about 0.95 of the area under ROC curves. The area of Tree-Lasso is about 0.84, that of LMM-Lasso is around 0.71. The area under ROC of MCP, SCAD, Lasso, LMM-BOLT, LTMLM and LMM-Select is 0.81, 0.81, 0.80, 0.57, 0.50 and 0.41 respectively.

C. PREDICTION OF Y

Figure 9 shows the Y results recovered. TgSLMM also yields the best result.[5]

[4] https://github.com/lebronlambert/TgSLMM

[5] The parameters that Figure 8 and Figure 9 used are just the same experimental setting in Section IV-C in our paper. n is 250; p is 500; k is 50; m is 10; d is 0.05; σ_e^2 is 0.001; σ_y^2 is 1; σ_ϵ^2 is 0.05; random seed is 0.



(a) Different magnitude of variance of response variables

(b) Different magnitude of noise

FIGURE 10. ROC curves for experiments with different parameters.

D. THE ROC CURVE OF SYNTHETIC EXPERIMENT

The Figure 10 shows the remaining images of ROC curves in synthetic experiments to compare our method with other existing methods by using the same parameters in our paper. For each configuration, the reported curve is drawn over five random seeds. And we can see that TgSLMM behaves almost always best.

ACKNOWLEDGMENT

(Hui Liu and Xiang Liu contributed equally to this work.) The authors would like to thank Haohan Wang from the University of Illinois Urbana–Champaign and Changpeng Lu from Rutgers University for early-stage efforts of this project. They also appreciate Zhou Fang for providing computational resource for part of the experiments.

REFERENCES

[1] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 1–10.

[2] L. Du and Y.-D. Shen, "Unsupervised feature selection with adaptive structure learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1–11.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Society: Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[4] Q. He and D.-Y. Lin, "A variable selection method for genome-wide association studies," *Bioinformatics*, vol. 27, no. 1, pp. 1–8, Jan. 2011.

[5] Q. Chen and S. Wang, "Variable selection for multiply-imputed data with application to dioxin exposure study," *Statist. Med.*, vol. 32, no. 21, pp. 3646–3659, Sep. 2013.

[6] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, and J. Ye, "Feafiner: Biomarker identification from medical data through feature generalization and selection," in *Proc. 19th ACM SIGKDD*, 2013, pp. 1–9.

[7] W. Astle and D. J. Balding, "Population structure and cryptic relatedness in genetic association studies," *Stat. Sci.*, vol. 24, no. 4, pp. 451–471, Nov. 2009.

[8] B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt, "A lasso multi-marker mixed model for association mapping with population structure correction," *Bioinformatics*, vol. 29, no. 2, pp. 206–214, Jan. 2013.

[9] H. Wang, B. Aragam, and E. P. Xing, "Variable selection in heterogeneous datasets: A truncated-rank sparse linear mixed model with applications to genome-wide association studies," *Methods*, vol. 145, pp. 2–9, Aug. 2018.

[10] H. Wang, X. Liu, Y. Xiao, M. Xu, and E. P. Xing, "Multiplex confounding factor correction for genomic association mapping with squared sparse linear mixed model," *Methods*, vol. 145, pp. 33–40, Aug. 2018.

[11] C. R. Henderson, "Best linear unbiased estimation and prediction under a selection model," *Biometrics*, vol. 31, no. 2, p. 423, Jun. 1975.

[12] H. Wang, T. Yue, J. Yang, W. Wu, and E. P. Xing, "Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies," *BMC Bioinf.*, vol. 20, no. S23, pp. 1–11, Dec. 2019.

[13] R. Dinga et al., "Controlling for effects of confounding variables on machine learning predictions," *bioRxiv*, 2020, doi: 10.1101/2020.08.17.255034.

[14] A. S. Hatoum, F. R. Wendt, M. Galimberti, R. Polimanti, B. Neale, H. R. Kranzler, J. Gelernter, H. J. Edenberg, and A. Agrawal, "Ancestry may confound genetic machine learning: Candidate-gene prediction of opioid use disorder as an example," *Drug Alcohol Dependence*, vol. 229, Dec. 2021, Art. no. 109115.

[15] X. Chen, S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing, "Graph-structured multi-task regression and an efficient optimization method for general fused lasso," 2010, *arXiv:1005.3579*.

[16] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping," *Ann. Appl. Statist.*, vol. 6, no. 3, pp. 1095–1117, Sep. 2012.

[17] W. Ye, X. Liu, T. Yue, and W. Wang, "A sparse graph-structured lasso mixed model for genetic association with confounding correction," 2017, *arXiv:1711.04162*.

[18] H. Wang, C. Lu, W. Wu, and E. P. Xing, "Graph-structured sparse mixed models for genetic association with confounding factors correction," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 298–302.

[19] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its Oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001.

[20] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, vol. 36, no. 4, p. 1509, Aug. 2008.

[21] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, Apr. 2010.

[22] H. Wang, B. J. Lengerich, B. Aragam, and E. P. Xing, "Precision lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data," *Bioinformatics*, vol. 35, no. 7, pp. 1181–1187, Apr. 2019.

[23] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[24] X. Tan, Y. Zhang, S. Tang, J. Shao, F. Wu, and Y. Zhuang, "Logistic tensor regression for classification," in *Proc. Int. Conf. Intell. Sci. Intell. Data Eng.*, 2012, pp. 573–581.

[25] T. J. Hayeck, N. A. Zaitlen, P.-R. Loh, B. Vilhjalmsson, S. Pollack, A. Gusev, J. Yang, G.-B. Chen, M. E. Goddard, P. M. Visscher, N. Patterson, and A. L. Price, "Mixed model with correction for case-control ascertainment increases association power," *Amer. J. Human Genet.*, vol. 96, no. 5, pp. 720–730, May 2015.

[26] H. Liu, M. Shao, and Y. Fu, "Consensus guided unsupervised feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[27] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet.*, vol. 2, no. 12, p. e190, 2006.

[28] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genet.*, vol. 38, no. 8, pp. 904–909, Aug. 2006.

[29] M. Goddard, "Genomic selection: Prediction of accuracy and maximisation of long term response," *Genetica*, vol. 136, no. 2, pp. 245–257, Jun. 2009.

[30] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin, "Variance component model to account for sample structure in genome-wide association studies," *Nature Genet.*, vol. 42, no. 4, pp. 348–354, Apr. 2010.

[31] H. Wang, B. Aragam, and E. P. Xing, "Trade-offs of linear mixed models in genome-wide association studies," *J. Comput. Biol.*, vol. 29, no. 3, pp. 233–242, Mar. 2022.

[32] J. Listgarten, C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman, "Improved linear mixed models for genome-wide association studies," *Nature Methods*, vol. 9, no. 6, pp. 525–526, Jun. 2012.

[33] P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price, "Efficient Bayesian mixed-model analysis increases association power in large cohorts," *Nature Genet.*, vol. 47, no. 3, pp. 284–290, Mar. 2015.

[34] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "FAST linear mixed models for genome-wide association studies," *Nature Methods*, vol. 8, no. 10, pp. 833–835, Oct. 2011.

[35] V. Segura, B. J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, and M. Nordborg, "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations," *Nature Genet.*, vol. 44, no. 7, pp. 825–830, Jul. 2012.

[36] M. Pirinen, P. Donnelly, and C. C. A. Spencer, "Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies," *Ann. Appl. Statist.*, vol. 7, no. 1, pp. 1–22, Mar. 2013.

[37] H. D. Bondell, A. Krishna, and S. K. Ghosh, "Joint variable selection for fixed and random effects in linear mixed-effects models," *Biometrics*, vol. 66, no. 4, pp. 1069–1077, Dec. 2010.

[38] Y. Fan and R. Li, "Variable selection in linear mixed effects models," *Ann. Statist.*, vol. 40, no. 4, p. 2043, Aug. 2012.

[39] A. Schwartzman, A. J. Schork, R. Zablocki, and W. K. Thompson, "A simple, consistent estimator of SNP heritability from genome-wide association studies," *Ann. Appl. Statist.*, vol. 13, no. 4, p. 2509, Dec. 2019.

[40] A. Pazokitoroudi, Y. Wu, K. S. Burch, K. Hou, A. Zhou, B. Pasaniuc, and S. Sankararaman, "Efficient variance components analysis across millions of genomes," *Nature Commun.*, vol. 11, no. 1, p. 4020, Aug. 2020.

[41] H. Wang, M. M. Vanyukov, E. P. Xing, and W. Wu, "Discovering weaker genetic associations guided by known associations," *BMC Med. Genomics*, vol. 13, no. S3, pp. 1–10, Feb. 2020.

[42] L. Jiang, Z. Zheng, H. Fang, and J. Yang, "A generalized linear mixed model association tool for biobank-scale data," *Nature Genet.*, vol. 53, no. 11, pp. 1616–1621, Nov. 2021.

[43] Y. Wu, K. S. Burch, A. Ganna, P. Pajukanta, B. Pasaniuc, and S. Sankararaman, "Fast estimation of genetic correlation for biobank-scale data," *Amer. J. Hum. Genet.*, vol. 109, no. 1, pp. 24–32, Jan. 2022.

[44] H. Wang, O. L. Lopez, W. Wu, and E. P. Xing, "Gene set priorization guided by regulatory networks with p-values through kernel mixed model," in *Proc. Int. Conf. Res. Comput. Mol. Biol.*, 2022, pp. 107–125.

[45] J. St-Pierre, K. Oualkacha, and S. R. Bhatnagar, "Efficient penalized generalized linear mixed models for variable selection and genetic risk prediction in high-dimensional data," *Bioinformatics*, vol. 39, no. 2, p. btad063, Feb. 2023.

[46] G. Kalantzis, "Methods for large-scale genome-wide association studies," Ph.D. dissertation, Dept. Statist., Univ. Oxford, Wellington Square, U.K., 2022.

[47] Z. Hou and A. Ochoa, "Genetic association models are robust to common population kinship estimation biases," *Genetics*, vol. 224, no. 1, p. iyad030, May 2023.

[48] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, "Smoothing proximal gradient method for general structured sparse regression," *Ann. Appl. Statist.*, vol. 6, no. 2, pp. 719–752, Jun. 2012.

[49] S. Atwell et al., "Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines," *Nature*, vol. 465, no. 7298, pp. 627–631, Jun. 2010.

[50] A. E. Anastasio, A. Platt, M. Horton, E. Grotewold, R. Scholl, J. O. Borevitz, M. Nordborg, and J. Bergelson, "Source verification of mis-identified arabidopsis thaliana accessions," *Plant J.*, vol. 67, no. 3, pp. 554–566, Aug. 2011.

[51] W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint, "Genome-wide genetic association of complex traits in heterogeneous stock mice," *Nature Genet.*, vol. 38, no. 8, pp. 879–887, Aug. 2006.

[52] B. Zhang et al., "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[53] P. Caramelli, R. Nitrini, R. Maranhão, A. C. G. Lourenço, M. C. Damasceno, C. Vinagre, and B. Caramelli, "Increased apolipoprotein B serum concentration in Alzheimer's disease," *Acta Neurologica Scandinavica*, vol. 100, no. 1, pp. 61–63, Jul. 1999.

[54] D. M. Williams, C. Finan, A. F. Schmidt, S. Burgess, and A. D. Hingorani, "Lipid lowering and Alzheimer disease risk: A Mendelian randomization study," *Ann. Neurol.*, vol. 87, no. 1, pp. 30–39, Jan. 2020.

[55] K. Wang, Y. Lu, D. F. Morrow, D. Xiao, and C. Xu, "Associations of ARHGAP26 polymorphisms with Alzheimer's disease and cardiovascular disease," *J. Mol. Neurosci.*, vol. 72, no. 5, pp. 1085–1097, May 2022.

[56] S. Lucken-Ardjomande Häsler, Y. Vallis, H. E. Jolin, A. N. McKenzie, and H. T. Mcmahon, "GRAF1a is a brain-specific protein promoting lipid droplet clustering and growth and enriched at lipid droplet junctions," *J. Cell Sci.*, vol. 127, no. 21, pp. 4602–4619, Jan. 2014.

[57] M. T. Flores-Dorantes, Y. E. Díaz-López, and R. Gutiérrez-Aguilar, "Environment and gene association with obesity and their impact on neurodegenerative and neurodevelopmental diseases," *Frontiers Neurosci.*, vol. 14, Aug. 2020, Art. no. 565326.

[58] S.-Y. Loke, P. T.-H. Wong, and W.-Y. Ong, "Global gene expression changes in the prefrontal cortex of rabbits with hypercholesterolemia and/or hypertension," *Neurochemistry Int.*, vol. 102, pp. 33–56, Jan. 2017.

[59] E. S. Luckett, M. Zielonka, A. Kordjani, J. Schaeverbeke, K. Adamczuk, S. De Meyer, K. Van Laere, P. Dupont, I. Cleynen, and R. Vandenberghe, "Longitudinal APOE4-and amyloid-dependent changes in the blood transcriptome in cognitively intact older adults," *Alzheimer's Res. Therapy*, vol. 15, no. 1, p. 121, Jul. 2023.

[60] Q. Y. Liu, R. R. Sooknanan, L. T. Malek, M. Ribecco-Lutkiewicz, J. X. Lei, H. Shen, B. Lach, P. R. Walker, J. Martin, and M. Sikorska, "Novel subtractive transcription-based amplification of mRNA (STAR) method and its application in search of rare and differentially expressed genes in AD brains," *BMC Genomics*, vol. 7, no. 1, pp. 1–6, Dec. 2006.

[61] W. Zhang, S. Thevapriya, P. J. Kim, W.-P. Yu, H. Shawn Je, E. King Tan, and L. Zeng, "Amyloid precursor protein regulates neurogenesis by antagonizing MIR-574-5P in the developing cerebral cortex," *Nature Commun.*, vol. 5, no. 1, p. 3330, Mar. 2014.

[62] M. P. Murphy and H. LeVine, "Alzheimer's disease and the amyloid-β peptide," *J. Alzheimer's Disease*, vol. 19, no. 1, pp. 311–323, 2010.

[63] D. van Abel, O. Michel, R. Veerhuis, M. Jacobs, M. van Dijk, and C. B. M. Oudejans, "Direct downregulation of CNTNAP2 by STOX1A is associated with Alzheimer's disease," *J. Alzheimer's Disease*, vol. 31, no. 4, pp. 793–800, Sep. 2012.

HUI LIU was born in China, in 1992. She received the bachelor's degree in traditional Chinese medicine from Binjiang College, First Clinical Medical College, Zhejiang Chinese Medical University, and the master's degree in public management from the School of Public Administration, Jiangxi Normal University. Her thesis defense was rated as excellent.

Her research interests include medical field, she utilizes clinical experimental data to model and incorporate modern medical techniques into traditional medical treatment approaches. Her focus is on observing the structural imbalances and disruptions in communication and information transfer at both the micro and macro levels of diseases, particularly related to endocrine hormone imbalances. She is concerned with pathological conditions resulting from disruptions in the physiological activities of organisms and aims to restore the body's equilibrium through pharmaceutical interventions. Additionally, in the field of bioinformatics, where she employs algorithmic models to analyze the relationships between animal genetic traits. She was awarded the First-Class Scholarship, the Merit Student Award, and the Outstanding Graduate Student Award, during the bachelor's degree.

XIANG LIU was born in Shandong, China, in 1995. He received the Bachelor of Engineering degree in communication engineering from Beijing University of Posts and Telecommunications, in 2018. He is currently pursuing the Ph.D. degree with the School of Computing, National University of Singapore. His major field of study has been focused on artificial intelligence systems, data science, and computer systems. He was once a Research Intern with Carnegie Mellon University, where he did research on bioinformatics. With these research experiences, he is conducting research that combines the knowledge of artificial intelligence and computer systems to facilitate database and data science applications. In terms of accolades, he received the national scholarship twice and has also been awarded a special scholarship. He has published one paper as the first author in the PSB 2019 Conference and received the Best Student Paper Award at the BIBM Conference. He has published papers in top-tier conferences and journals, such as CVPR, VTC, and *Methods*. Additionally, he has submitted one patent application related to using deep learning to assist the visually impaired.

JING DIAO was born in Shandong, China, in 1995. She received the bachelor's degree in stomatology from Peking University, in 2019. She is currently pursuing the M.D. degree with Peking University School and Hospital of Stomatology. Her major field of study has been focused on biomarkers of oral diseases.

She has published one article as the first author in *Frontiers in Cellular and Infection Microbiology*. Her research interests include utilizing multi-omics approaches, including microbiome and metabolomics, to discover biomarkers for oral diseases. When dealing with large datasets, she employs artificial intelligence techniques, such as machine learning and bioinformatics tools for modeling, facilitating the identification of disease biomarkers.

WENTING YE was born in Fuzhou, China, in 1996. He received the Bachelor of Engineering degree in network engineering from Beijing University of Posts and Telecommunications, in 2018, and the master's degree in computational data science from Carnegie Mellon University, Pittsburgh, PA, USA, in 2019. His major field of study has been focused on machine learning and data science.

He has accumulated a wealth of experience in both academia and industry. He was once a Research Intern with Carnegie Mellon University, where he focused on bioinformatics. Starting his career as a Machine Learning Intern with Bytedance AI Lab, Beijing, he later joined Waymo, as a Scene Understanding Intern, Mountain View, CA, USA. He then, was an Applied Scientist with Amazon.com Services, Inc., Seattle, WA, USA, where he led the scientific design for private brand discovery and developed a multilingual transformer-based model. He is currently a Senior Machine Learning Engineer with ByteDance Inc., Bellevue, WA, USA, where he involved on fundamental algorithms for user modeling and ads ranking. He has published one paper as the first author in the ACM SIGIR 2023 Conference. Additionally, he has submitted one patent application related to multi-task learning in user modeling.

Mr. Ye served on the Program Committee for several major conferences, such as ACL, SIGIR, NeurIPS, and EMNLP. In terms of accolades, he received the national scholarships three times.

XUELING LIU was born in Zhejiang, China, in 1996. She received the bachelor's degree in business administration from Beijing University of Posts and Telecommunications, in 2018, and the master's degree in biomedical informatics from the National University of Singapore, in 2022. Her major field of study has been focused on bioinformatics and biomedical informatics.

Her career began in the field of biomedical science and she is currently exploring how to use computer tools more effectively to extract biomedical information to assist in medical research. In the near future, she may join the National University of Singapore, as a Research Assistant. During her academic journey, she received several school scholarships.

DEHUI WEI (Graduate Student Member, IEEE) received the B.S. degree in computer science and technology from Hunan University, Changsha, China, in 2019. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT). Her research interests include network transmission control, cloud computing, and data processing. During the B.S. degree, she was awarded the Outstanding Graduate Award.