HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2018 April 06.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2017 November ; 2017: 431–438. doi:10.1109/BIBM.2017.8217687.

Variable Selection in Heterogeneous Datasets: A Truncated-rank Sparse Linear Mixed Model with Applications to Genome-wide Association Studies

Haohan Wang,

Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Bryon Aragam, and

Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Eric P. Xing

Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Abstract

A fundamental and important challenge in modern datasets of ever increasing dimensionality is variable selection, which has taken on renewed interest recently due to the growth of biological and medical datasets with complex, non-i.i.d. structures. Naïvely applying classical variable selection methods such as the Lasso to such datasets may lead to a large number of false discoveries. Motivated by genome-wide association studies in genetics, we study the problem of variable selection for datasets arising from multiple subpopulations, when this underlying population structure is unknown to the researcher. We propose a unified framework for sparse variable selection that adaptively corrects for population structure via a low-rank linear mixed model. Most importantly, the proposed method does not require prior knowledge of individual relationships in the data and adaptively selects a covariance structure of the correct complexity. Through extensive experiments, we illustrate the effectiveness of this framework over existing methods. Further, we test our method on three different genomic datasets from plants, mice, and humans, and discuss the knowledge we discover with our model.

Keywords

Variable Selection; Confounding Correction; Computational Ge-nomics; Sparsity; Linear Mixed Model

KDD'17, Halifax, Nova Scotia, Canada

CCS CONCEPTS

Information systems → Data mining; Computing methodologies → Supervised learning; Applied computing → *Genetics*; *Computational genomics*

1 INTRODUCTION

Increasingly, modern datasets are derived from multiple sources such as different experiments, different databases, or different populations. In combining such heterogeneous datasets, one of the most fundamental assumptions in statistics and machine learning is violated: That observations are independent of one another. When a dataset arises from multiple sources, dependencies are introduced between observations from similar batches, regions, populations, etc. As a result, classical methods breakdown and novel procedures that can handle heterogeneous datasets and correlated observations are becoming more and more important.

In this paper, we focus on the important problem of variable selection in non-i.i.d. settings with possibly dependent observations. In addition to the aforementioned complications in analyzing datasets arising from multiple sources, the rapid increase in the dimensionality of data continues to hasten the need for reliable variable selection procedures to reduce this dimensionality. This issue is especially salient in genomics applications in which datasets routinely contain hundreds of thousands of genetic markers coming from different populations. For example, to discover genomic associations for a certain disease, genetic data from patients is often collected from different hospitals. As a result, data from the case and control groups can be confounded with variables such as the hospital, clinical trial, city, or even country. Another common source of sample dependence is family relatedness and population ancestry between individuals [2].

Unfortunately, in many applications information on the origin of different observations is lost either through data compression or experimental necessity. For example, for privacy reasons, it may be necessary to anonymize datasets thereby obfuscating the relationship between different observations. As a result, the data becomes confounded and attempts to learn associations via existing variable selection procedures are doomed to fail [15]. In seeking to discover information from such rich datasets when we do not have this important information, it becomes necessary to *deconfound* our models in order to implicitly account for this.

Example 1.1

Figure 1 gives a concrete example of how multi-sourced data introduces confounding factors. In this toy example, we have collected seven samples of data from a diseased case group and another seven samples from the healthy control group to study 25 genetic variables (in this case, nucleotides). Simple statistics will tell us that the 11^{th} variable has the strongest association with the disease in question. However, suppose that a closer look at the data reveals that these samples originate from two different populations. As a result, the association between disease status and the 11^{th} nucleotide becomes confounded by

population membership. In this toy dataset, we cannot avoid this false discovery unless we know which population each sample originates from.

Existing solutions rely on traditional hypothesis testing after a dedicated confounding correction step, usually resulting in suboptimal performance [18, 44]. In contrast, state-of-the-art variable selection methods usually assume that the data comes from a single distribution, leading to reduced performance when applied to multi-source data.

We directly address the problem of variable selection with heterogeneous data by integrating state-of-the-art confounding correction methods and variable selection methods, thereby introducing a general methodology for addressing this issue. Our main contributions are the following:

- We propose a general sparse variable selection framework that takes into account possibly heterogeneous datasets by implicitly correcting for confounders via linear mixed models,
- We improve this framework by introducing an adaptive procedure for automatically selecting a low-rank approximation in the linear mixed model,
- We apply our model to three distinct genomic datasets in order to illustrate the effectiveness of the method on a real application.

The remainder of the paper is organized as follows: We begin by briefly summarizing the previous approaches to sparse variable selection and confounding correction respectively in Section 2. Then in Section 3 we propose a general framework to accomplish both variable selection and confounder correction simultaneously. In Section 4, we show how using a low-rank approximation can be used to improve this framework, thereby introducing the Truncated-rank Sparse Linear Mixed Model. The performance of our model over traditional methods is validated with synthetic datasets in Section 5. Finally, in Section 6, we validate our method on three real-world genomic datasets.

2 RELATED WORK

Variable selection is a fundamental problem in knowledge discovery and has attracted significant attention from the machine learning and statistical communities. The basic idea is to reduce the dimensionality of a large dataset by selecting a subset of representative features without substantial loss of information [5, 7, 24, 39]. This problem has attracted substantial attention in the so-called *high-dimensional* regime, where it is typically assumed that only a small subset of features are relevant to a response. In order to identify this subset, arguably the most popular method is ℓ_1-norm regularization (i.e. *Lasso* regression [33]). More recently, nonconvex regularizers have been introduced to overcome the limitations of Lasso [8]. Examples include the Smoothly Clipped Absolute Deviation (SCAD) [8] and the Minimax Concave Penalty (MCP) [41]. These methods overcome many of the aforementioned limitations at the cost of introducing nonconvexity in the optimization problem; a recent review of these methods can be found in [42]. In applications, variable selection is broadly used to extract variables that are interpretable or potentially causal [16, 38], especially in biology [12] and medicine [6, 43].

When the data is non-i.i.d., such as when it arises from distinct subpopulations, two popular approaches for addressing this are principal component analysis [27, 29] and linear mixed models [10, 15]. Mixed models first rose to prominence in the animal breeding literature, where they were used to correct for kinship and family structure [13]. Interest in these methods has surged recently given improvements that allow their application to human-scale genome data [20, 28, 32]. These methods, however, ultimately rely on classical hypothesis testing procedures for variable selection after confounding correction. Finally, a recent line of work has sought to combine the advantages of linear mixed models with sparse variable selection [4, 9, 31, 37].

3 SPARSE VARIABLE SELECTION WITH CONFOUNDING CORRECTION

Before we introduce our general framework, we first revisit the classical linear mixed model [22].

3.1 Linear Mixed Model

The linear mixed model (LMM) is an extension of the standard linear regression model that explicitly describes the relationship between a response variable and explanatory variables incorporating an extra, random term to account for confounding factors. As a consequence, a mixed-effects model consists of two parts: 1) Fixed effects corresponding to the conventional linear regression covariates, and 2) Random effects that account for confounding factors.

Formally, suppose we have n samples, with response variable $y = (y_1, y_2, \ldots y_n)$ and known explanatory variables $X = (x_1, x_2, \ldots x_n)$.

For each $i = 1, 2, \ldots, n$, we have $x_i = (x_{i,1}, x_{i,2}, \ldots x_{i,p}$, i.e., X is of the size $n \times p$. The standard linear regression model asserts $y = X\beta + \varepsilon$, where β is an unknown parameter vector and $\varepsilon \sim N(0, \sigma_e^2 I)$. In the linear mixed model, we add a second term $Z\mu$ to model confounders:

$$y = X\beta + Z\mu + \varepsilon, \quad (1)$$

Here, Z is a known $n \times t$ matrix of *random effects* and μ is a random variable. Intuitively, the product $Z\mu$ models the covariance between the observations y_i. This can be made explicit by further assuming that $\mu \sim N(0, \sigma_g^2 I)$, in which case we have

$$y \sim N(X\beta, \sigma_g^2 K + \sigma_e^2 I) \quad (2)$$

where $K = ZZ^T$. Here, K explicitly represents the covariance between the observations (up to measurement error $\sigma_e^2 I$): If $K = 0$, then each y_i is uncorrelated with the rest of the observations and we recover the usual linear regression model. When $K \neq 0$, we have a nontrivial linear mixed model. As K is required to be known, early applications of LMMs also assumed that K was known in advance [13]. Unfortunately, in many cases (including

genetic studies), this information is not known in advance. In these cases, a common convention is to estimate K from the available explanatory variables. As we shall see in Section 4, finding a good approximation to K is crucial to obtaining good results in variable selection.

3.2 Sparsity Regularized Linear Mixed Model

For high-dimensional models with $p \gg n$, it is often of interest to regularize the resulting model to select out important variables and simplify its interpretation. This can easily be achieved by introducing sparsity-inducing priors to the posterior distribution. For example, [31] introduces the Laplace prior, which leads to a ℓ_1 regularized linear mixed model as following:

$$p(\beta, \sigma_g, \sigma_e | y, X, K) \propto N(y | X\beta, \sigma_g^2 K + \sigma_e^2 I) e^{-\lambda |\beta|}$$

We call the result the *sparse linear mixed model*, or SLMM for short.

This choice of prior—which corresponds to the well-known Lasso when only fixed effects are considered—is well-known to suffer from limitations in variable selection [8]. In this paper, we extend this SLMM-Lasso model to more advanced regularization schemes such as the MCP and SCAD, which we call the SLMM-SCAD and SLMM-MCP, respectively. For simplicity, we will use $f(\beta)$ to denote a general regularizer, yielding the following general posterior:

$$p(\beta, \sigma_g, \sigma_e | y, X, K) \propto N(y | X\beta, \sigma_g^2 K + \sigma_e^2 I) e^{-f(\beta)}. \quad (3)$$

This allows us to combine the (independently) well-studied advantages of the linear mixed model for confounding correction with those of high-dimensional regression for variable selection.

4 TRUNCATED-RANK SPARSE LINEAR MIXED MODEL

Despite their successes, the main drawback of the aforementioned mixed model approaches is the estimation of K from the data X. In this section, we propose an adaptive, low-rank approximation for K in order to more accurately model latent population structure.

4.1 Motivation

Even though K is assumed to be known in LMMs, we have already noted that in practice K is often unknown. Thus, to emphasize the distinction between the true, *unknown* covariance K and an estimate based on data, we let $\widetilde{K} = \widetilde{K}(X)$ denote such an estimate. Substituting $\widetilde{K}$ for K in (3), the posterior then becomes:

$$p(\beta, \sigma_g, \sigma_e | y, X, \widetilde{K}) \propto N(y | X\beta, \sigma_g^2 \widetilde{K} + \sigma_e^2 I) e^{-f(\beta)}. \quad (4)$$

By far the most common approximation used in practice is $\widetilde{K} = XX^T$ [13]. Under this approximation, equation 1 becomes

$$y = X\beta + X\mu + \varepsilon = X(\beta + \mu) + \varepsilon$$

where $\mu \sim N(0, \sigma_\mu^2)$. As our goal is the estimation of β, this evidently makes distinguishing β and μ difficult.

This approximation was originally motivated as a way to use the observed variables X as a surrogate to model the relationship between the observations y. The hope is that the values in X might cluster conveniently according to different batches, regions, or populations, which are the presumed sources of confounding. One straightforward observation is that such sources of confounding typically have a much lower dimensionality than the total number of samples in the data. As a result, we expect that K will have a low-rank structure which we can and should exploit. Unfortunately, the matrix XX^T will not, in general, be low-rank—in fact, it can be *full rank*, with $\text{rank}(XX^T) = n$. To correct for this, we propose the Truncated-rank Sparse Linear Mixed Model (TrSLMM).

4.2 Truncated-rank Sparse Linear Mixed Model

Instead of choosing $\widetilde{K} = XX^T$ as our approximation, we seek a low-rank approximation to the true covariance K. Let $\Gamma := XX^T$ and $\Gamma = U\Lambda V^T$ be the SVD of Γ. Define Λ_s to be the diagonal matrix such that $(\Lambda_s)_{jj} = \Lambda_{jj}$ for $j \leq s$ and $(\Lambda_s)_{jj} = 0$ otherwise (assuming values of Λ are in decreasing order). Then a natural choice for $\widetilde{K}$ is $\Gamma_s := U\Lambda_s V^T$ for some $0 < s < n$, i.e. the best s-rank approximation to Γ.

Selection of s. Of course, we have simply replaced the problem of estimating K with that of estimating an optimal rank s from the data. Fortunately, the latter can be done efficiently. To motivate the selection of s, we first investigate the distribution of Λ under different population structures. Let G denote the number of subpopulations or distributions used to generate the data. Figure 2 shows a plot of normalized Λ for 100 data samples for $G = 1$, 5, 20, 100. We can clearly see that in the middle two cases ($G = 5$ and $G = 20$), the singular values exhibit some interesting patterns: Instead of decaying smoothly (as for $G = 1$ and $G = 100$), there are a few dominant singular values and more small singular values following a steep drop-off. This confirms our intuition of a latent, approximately low-rank structure within Γ.

Based on this observation, we introduce a clean solution to truncate Λ: We can directly screen out the top, dominant singular values by selecting the top s values Λ_j for which

$$\frac{\Lambda_j - \Lambda_{j+1}}{\Lambda_0} > \frac{1}{n}$$

where n is the number of samples. In particular, the number of selected singular values s satisfies $(\Lambda_s - \Lambda_{s+1})/\Lambda_0 > 1/n$ and $(\Lambda_{s-1} - \Lambda_s)/\Lambda_0 \leq 1/n$.

Then, we have:

$$(\Lambda_s)_{jj} = \begin{cases} \Lambda_{jj} & \text{if } j \leq s \\ 0 & \text{otherwise} \end{cases}$$

and finally:

$$\widetilde{K} = \Gamma_s = U\Lambda_s V^T$$

A similar low-rank approximation idea has been used previously [15, 30], however, these procedures require specifying unknown hyperparameters, even when replaced by sparse PCA [45] or Bayesian *K*-means [19]. Another approach is to fit every possible low-rank Λ_s sequentially and selecting the best configuration of singular values based on a pre-determined criteria [14], which is $O(n)$ slower than our method and most importantly does not scale for modern human genome datasets.

4.3 Parameter Learning

In order to infer the parameters $\{\beta, \sigma_g, \sigma_e\}$, we break the problem into two steps: 1) Confounder correction, where we solve for σ_g and σ_e; and 2) Sparse variable selection, where we solve for β in Equation 4.

Confounder Correction—Following the empirical results in [15], we first estimate the variance term with:

$$p(\sigma_g, \sigma_e | y, \widetilde{K}) \propto N(y - \bar{y} | 0, \sigma_g^2 \widetilde{K} + \sigma_e^2 I) \quad (5)$$

where $\bar{y}$ is the empirical mean of y. We then solve Equation 5 for σ_g and σ_e, where we can adopt the trick of introducing $\delta = \frac{\sigma_e^2}{\sigma_g^2}$ to replace σ_g^2 for more efficient optimization [20].

Finally, we can then correct the confounding factors by rotating the original data:

$$X' = (\text{diag}(\Gamma_s) + \delta I)^{-\frac{1}{2}} V^T X$$

$$y' = (\text{diag}(\Gamma_s) + \delta I)^{-\frac{1}{2}} V^T y$$

where $\widetilde{K} = U\Gamma_s V^T$ is the singular value decomposition, which has already been computed to determine s.

Sparse Variable Selection—After rotating the data to produce X' and y', we have a standard variable selection task at hand [31]. Thus, maximizing the posterior in Equation 4

becomes equivalent to solving a variable selection problem with X' and y'. Note that unlike vanilla linear regression, which would be unchanged by rotations, the introduction of the random effects $Z\mu$ in (2) violates this rotation-invariance property.

For different choices of regularizer $f(\beta)$, we can then solve the following regularized linear regression problem:

$$\arg\min_{\beta} \|y' - X'\beta\|_2^2 + f(\beta)$$

where standard optimization techniques can be adopted. In our experiments, we use proximal gradient descent [26].

5 SYNTHETIC EXPERIMENTS

In this section, we evaluate the performance of our proposed Truncated-rank Sparse Linear Mixed Model (TrSLMM) against vanilla sparse linear mixed models as well as classical sparse variable selection methods.

5.1 Data Generation

We first simulate observed covariates coming from G different populations. We use c_g to denote the centroid of the gth population, $g = 1, \ldots, G$. First, we generate the centroids c_g and from each centroid, we generate explanatory variables from a multivariate Gaussian distribution as follows:

$$x_{ij} = N(c_g, \sigma_e^2 I)$$

where x_{ij} denotes the ith data from gth distribution.

We then generate an intermediate response r from X from the usual linear regression model:

$$r = X\beta + \varepsilon. \quad (6)$$

Here β is a sparse vector indicating which variables in X influences the outcome r and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

Note that the components)of r are uncorrelated—in order to simulate a scenario with correlated observations, we introduce a covariance matrix to simulate correlations between the y_i. Thus, we generate the final response y as follows:

$$y \sim N(r, \sigma_y^2 M) \quad (7)$$

where M is the covariance between observations and σ_y^2 is a scalar that controls the magnitude of the variance. Letting C be the matrix formed by stacking the centroids c_g, we choose $M = CC^T$. This has the desired effect of making observations from the same group g more correlated.

Figure 3 shows an illustration of the synthetic data generated according to our data generation recipe. The x-axis and y-axis show the first two principle components of X, and the z-axis shows the final response variable y. Evidently, both X and y fall into five clusters.

5.2 Experimental Results in Variable Selection

We use the following parameters in our simulations:

- n the number of data samples. The default is 1000.
- p the number of explanatory variables. The default is 5000.
- d the percentage of explanatory variables to be active variables. The default is 0.05.
- G the number of distributions the data originate from. The default is 10.
- σ_e the magnitude of covariance of explanatory variables. The default is 0.1.
- σ_r the magnitude of covariance of response variables. The default is 1.

The results are shown as ROC curves in Figure 4. In general, across all the parameter settings tested, we see that the proposed Truncated-rank Sparse Linear Mixed Model outperforms the other methods. Unsurprisingly, the Sparse Linear Mixed Model outperforms traditional sparse variable selection methods, which was completely ineffective in this experiment. This illustrates how methods that do not account for possible sources of confounding can drastically underperform when the assumption that observations are independent is violated.

As the various parameters are changed, we observe some expected patterns. For example, in Figure 4(a), as n increases, and in Figure 4(b) as p decreases, the ratio of $\frac{p}{n}$ gets smaller and the performance gets better. As we increase the proportion of nonzero coefficients in β, the number of distributions, or the variance of response variable y, the problem becomes more challenging. In almost all of these cases, however, the TrSLMM-based methods show improved performance. As an example where the SLMM methods are comparable when $G = 2$ SLMM-MCP and SLMM-SCAD behave better than TrSLMM-Lasso, but even they remain slightly inferior to TrSLMM-MCP and TrSLMM-SCAD. Traditional variable selection methods, for the most part, show the same behavior as these parameters are manipulated—this suggests that the fluctuations we observe in the other methods are due to the different strategies by which confounding is corrected.

5.3 Prediction of True Effect Sizes

Figure 5 shows the averaged mean squared error in estimating the effect sizes β and its standard error over five runs for different settings when we adjust the percentage of causal

variables d on synthetic data. Interestingly, we can see that TrSLMM-Lasso behave the best in estimating β, while SLMM-Lasso closely follows-up. Traditional sparse variable selection methods (Lasso, SCAD, MCP) behave worse than these two methods, but mostly better than other TrSLMM and SLMM based methods.

5.4 Running Time

After confounding correction, we observed that the final sparse variable selection step converged faster. Across all the configurations of synthetic experiments, in comparison to the vanilla sparse variable selection methods, TrSLMM-Lasso, TrSLMM-SCAD, and TrSLMM-MCP only required 49%, 38%, and 29%, respectively, of the time needed for the Lasso, SCAD, and MCP, respectively, to converge on average. SLMM-Lasso, SLMM-SCAD, SLMM-MCP were slightly faster, and only required 28%, 38%, 37% of the time needed on average. While not necessarily faster overall, this is an interesting observation and confirms previous theoretical work suggesting that variable selection is faster and easier for uncorrelated variables.

6 REAL GENOME DATA EXPERIMENTS

In order to evaluate the TrSLMM framework in a practical setting, we tested our model on three datasets coming from genomics studies. To provide a clearer evaluation, we tested our method on datasets from three different species. We then evaluate our discovered knowledge with some of the published results in relevant literature to show the reliability of our methods compared with existing approaches. Finally, we report our discovered associations.

6.1 Data Sets

6.1.1 Arabidopsis thaliana—The Arabidopsis thaliana dataset we obtained is a collection of around 200 plants, each with around 215,000 genetic variables [1]. We study the association between these genetic variables and a set of observed traits. These plants were collected from 27 different countries in Europe and Asia, so that geographic origin serves as a potential confounding factor. For example, different sunlight conditions in different regions may affect the observed traits of these plants. We tested the genetic associations between genetic variables with 44 different traits such as *days to germination*, *days to flowering*, *lesioning* etc.

6.1.2 Heterogeneous Stock Mice—The heterogeneous stock mice dataset contains measurements from around 700 mice, with 100,000 genetic variables [34]. These mice were raised in cages by four generations over a two-year period. In total, the mice come from 85 distinct families. The obvious confounding variable here is genetic inheritance due to family relationships. We studied the association between the genetic variables and a set of 28 response variables that could possibility be affected by inheritance. These 28 response variables fall into three different categories, relating to the glucose level, insulin level and immunity respectively.

6.1.3 Human Alzheimer's Disease—We use the late-onset Alzheimer's Disease data provided by Harvard Brain Tissue Resource Center and Merck Research Laboratories [40].

It consists of measurements from 540 patients with 500,000 genetic variables. We tested the association between these genetic variables and a binary response corresponding to a patient's disease status of Alzheimer's disease.

6.2 Ground Truth for Evaluation

To evaluate the performance of TrSLMM, we compared the results with genetic variables that have been reported in the genetics literature to be associated with the response variables of interest. Although new associations may yet be discovered by the genetics community, Arabidopsis thaliana and mice have been studied for over a decade, and the scientific community has reached a general consensus regarding these species. For Arabidopsis thaliana, we use the validated knowledge of the genetic associations reported in [3] to evaluate the performance of these models. For heterogeneous stock mice, the validated gold standard genetic variables were collected from the Mouse Genome Informatics database.[1] Although Alzheimer's disease is a very active area of research, there is no consensus gold standard available. Instead, we listed the genetic variables identified by one of our proposed model (TrSLMM-MCP) and verified the top genetic variables by searching the relevant literature. Additionally, since the genetic cause of Alzheimer's disease is still an open research area, we reported the genetic variables we identified for the benefit of domain experts.

6.3 Selected Groups

We first validate the success of our truncated-rank approaches to identify the truly confounding factors from distributions of eigenvalues. Figure 6 shows the distribution of eigenvalues of XX^T. A naïve Linear Mixed Model will correct the confounding factors with all these eigenvalues, resulting in an over-correction. In contrast, Truncated-rank Sparse Linear Mixed Model only identifies the ones that are likely to be confounding sources. As Figure 6 shows, TrSLMM conveniently identifies 27 data origins for Arabidopsis thaliana, while these 200 plants are in fact collected from 27 countries. TrSLMM identifies 65 sources for mice data, while these mice are from 85 different families. Although TrSLMM didn't pinpoint every confounding factor exactly, the number of confounding factors is much closer compared to vanilla sparse variable selection methods (only one) and vanilla SLMM methods (number of samples by construction). On the human Alzheimer's Disease, there is no consensus number of data sources available to check the correctness of TrSLMM's selection, but the distribution seems to indicate that there are only a few confounding sources.

6.4 Numerical Results

Since we have access to a validated gold standard in two out of the three datasets, Figure 7 and Figure 8 illustrate the area under the ROC curve for each response variables (observed trait) for Arabidopsis thaliana and Mice, respectively. The responses are ordered such that the leftmost variables are those for which our TrSLMM model outperform the others. Because discovering associations in genetic datasets is an extremely challenging task, many

[1] http://www.informatics.jax.org/

of these methods fail to discover useful variables. It is worth emphasizing that the discovery of even a few highly associated variants can be significant in practice. Overall, TrSLMM methods managed to outperform the other methods for almost 60% of response variables. TrSLMM-MCP and TrSLMM-SCAD behave similarly, as previously observed in the synthetic data experiments.

For Arabidopsis thaliana, TrSLMM based models behave as the best one on 56.8% of the traits. Since not all of the traits in our dataset are expected to be confounded, it is not surprising that in some cases traditional methods perform well. Without confounding, one expects methods that are optimized for i.i.d. data to perform best (e.g. Lasso, SCAD, MCP). For example, traits with **GH** in the name mean that the corresponding traits were measured in a greenhouse, where conditions are strictly controlled and potential confounding effects introduced by different regions are minimized. As Figure 7 shows, traditional sparse variable selection methods almost gain the most advantage over greenhouse traits.

For Heterogeneous Stock Mice, TrSLMM based models behave as the best one on 57.4% of the traits. The results are interesting: The left side of the figure mostly consists of traits regarding the amount of glucose and insulin in the mice, while the right hand side of the figure mostly consists of traits related to immunity. This raises the interesting question of whether or not immune levels in stock mice are largely independent of family origin.

Most importantly, our proposed model is at least as good as other SLMM based methods, and sometimes significantly better when confounding is present. This gain in performance comes with no extra parameters and no extra computation, except for one computationally trivial step of screening singular values.

6.5 Knowledge Discovered and Causality Analysis

Finally, we proceed to the Human Alzheimer's Disease dataset. Because Alzheimer's Disease has not been studied as extensively as plants and mice, there is no authentic golden standard to evaluate the performances. Here, we report the top 99 genetic variables our model discovered in Table 1 to foster relevant research.

Due to space limitations, we briefly justify only a few of the most important genetic variables here. The 4th discovered variable is associated with *AOPE* gene, which is the gene that is prominently believed to be cause Alzheimer's disease [21]. The 5th discovered variable is associated with *COL1A1* gene, which is associated with *APOE* [25]. The 6th one is *WFDC1*, which is reported to be associated with Alzheimer's disease [23]. The 9th one is associated with *GALNTL4*, which is reported to be related with Alzheimer's disease [11].

7 CONCLUSIONS

In this paper, we aim to solve a critical challenge in variable selection when the data is not i.i.d. and does not come from the same distribution. Due to confounding, traditional variable selection procedures tend to select variables that are not relevant. When the sources of confounding are known and can be controlled for, linear mixed models have long been used to make such corrections. The use of LMMs to *implicitly* correct for confounding that is not

explicitly known to an analyst is a recent development and a very active area of research. This type of situation occurs frequently in genomics applications where confounding arises due to population stratification, batch effects, and family relationships.

To overcome this problem, we introduced a general framework for sparse variable selection from heterogeneous datasets. The procedure consists of a confounding correction step via linear mixed models followed up by sparse variable selection. We have shown that state-of-the-art variable selection methods such as SCAD and MCP can be easily plugged into this procedure. Further, we showed that the traditional linear mixed model can easily fall into the trap of utilizing too much information, resulting in an over-correction. To correct for this, we introduce a Truncated-rank Sparse Linear Mixed Model that effectively and automatically identifies the sources of confounding factors. Most importantly, we proposed a data-driven, adaptive procedure to automatically identify confounding sources from the spectrum of the kinship matrix without prior knowledge. Through extensive experiments, we exhibited how TrSLMM has a clear advantage over existing methods in synthetic experiments and real genome datasets across three different species: plant (Arabidopsis thaliana), mice, and human.

In future work, we plan to explore more complex structured problems with our proposed framework to select variables for response variables that are dependent [17] or for explanatory variables that are correlated [35]. Further, we plan to integrate our method into the popular genomic research toolbox GenAMap [36].

References

1. Anastasio, Alison E., Platt, Alexander, Horton, Matthew, Grotewold, Erich, Scholl, Randy, Borevitz, Justin O., Nordborg, Magnus, Bergelson, Joy. Source verification of mis-identified Arabidopsis thaliana accessions. The Plant Journal. 2011; 67(3):554–566. 2011. [PubMed: 21481029]
2. Astle, William, Balding, David J. Population structure and cryptic relatedness in genetic association studies. Statist Sci. 2009; 2009:451–471.
3. Atwell, Susanna, Huang, Yu S., Vilhjálmsson, Bjarni J., Willems, Glenda, Horton, Matthew, Li, Yan, Meng, Dazhe, Platt, Alexander, Tarone, Aaron M., Hu, Tina T., et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010; 465(7298):627–631. 2010. [PubMed: 20336072]
4. Bondell, Howard D., Krishna, Arun, Ghosh, Sujit K. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. Biometrics. 2010; 66(4):1069–1077. 2010. [PubMed: 20163404]
5. Cai, Deng, Zhang, Chiyuan, He, Xiaofei. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2010. Unsupervised feature selection for multi-cluster data; p. 333-342.
6. Chen, Qixuan, Wang, Sijian. Variable selection for multiply-imputed data with application to dioxin exposure study. Statistics in medicine. 2013; 32(21):3646–3659. 2013. [PubMed: 23526243]
7. Du, Liang, Shen, Yi-Dong. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2015. Unsupervised feature selection with adaptive structure learning; p. 209-218.
8. Fan, Jianqing, Li, Runze. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association. 2001; 96(456):1348–1360. 2001.
9. Fan, Yingying, Li, Runze. Variable selection in linear mixed effects models. Annals of statistics. 2012; 40:4.

10. Goddard, Mike. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136(2):245–257. 2009. [PubMed: 18704696]
11. Harold, Denise, Abraham, Richard, Hollingworth, Paul, Sims, Rebecca, Gerrish, Amy, Hamshere, Marian L., Pahwa, Jaspreet Singh, Moskvina, Valentina, Dowzell, Kimberley, Williams, Amy, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nature genetics. 2009; 41(10):1088–1093. 2009. [PubMed: 19734902]
12. He, Qianchuan, Lin, Dan-Yu. A variable selection method for genome-wide association studies. Bioinformatics. 2011; 27(1):1–8. 2011. [PubMed: 21036813]
13. Henderson, Charles R. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975; 1975:423–447.
14. Hoffman, Gabriel E. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS One. 2013; 8(10):e75707. 2013. [PubMed: 24204578]
15. Kang, Hyun Min, Sul, Jae Hoon, Service, Susan K., Zaitlen, Noah A., Kong, Sit-yee, Freimer, Nelson B., Sabatti, Chiara, Eskin, Eleazar, et al. Variance component model to account for sample structure in genome-wide association studies. Nature genetics. 2010; 42(4):348–354. 2010. [PubMed: 20208533]
16. Kim, Been, Shah, Julie A., Doshi-Velez, Finale. Mind the gap: A generative approach to interpretable feature selection and extraction. Advances in Neural Information Processing Systems. 2015:2260–2268.
17. Kim, Seyoung, Xing, Eric P. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. The Annals of Applied Statistics. 2012; 2012:1095–1117.
18. Korte, Arthur, Farlow, Ashley. The advantages and limitations of trait analysis with GWAS: a review. Plant methods. 2013; 9:1. [PubMed: 23286457]
19. Kulis, Brian, Jordan, Michael I. Revisiting k-means: New algorithms via Bayesian nonparametrics. arXiv preprint arXiv:1111.0352. 2011 2011.
20. Lippert, Christoph, Listgarten, Jennifer, Liu, Ying, Kadie, Carl M., Davidson, Robert I., Heckerman, David. FaST linear mixed models for genome-wide association studies. Nature methods. 2011; 8(10):833–835. 2011. [PubMed: 21892150]
21. Liu, Chia-Chan, Kanekiyo, Takahisa, Xu, Huaxi, Bu, Guojun. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nature Reviews Neurology. 2013; 9(2):106–118. 2013. [PubMed: 23296339]
22. McCulloch, Charles E., Neuhaus, John M. Generalized linear mixed models. Wiley Online Library; 2001.
23. Miller, Jeremy A., Woltjer, Randall L., Goodenbour, Jeff M., Horvath, Steve, Geschwind, Daniel H. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. Genome medicine. 2013; 5(5):48. 2013. [PubMed: 23705665]
24. Nie, Feiping, Huang, Heng, Cai, Xiao, Ding, Chris H. Efficient and robust feature selection via joint —fi2, 1-norms minimization. Advances in neural information processing systems. 2010:1813–1821.
25. Oue, Naohide, Hamai, Yoichi, Mitani, Yoshitsugu, Matsumura, Shunji, Oshimo, Yasuhiro, Aung, Phyu Phyu, Kuraoka, Kazuya, Nakayama, Hirofumi, Yasui, Wataru. Gene expression profile of gastric carcinoma. Cancer research. 2004; 64(7):2397–2405. 2004. [PubMed: 15059891]
26. Parikh, Neal, Boyd, Stephen, et al. Proximal algorithms. Foundations and Trends® in Optimization. 2014; 1(3):127–239. 2014.
27. Patterson, Nick, Price, Alkes L., Reich, David. Population structure and eigenanalysis. PLoS genet. 2006; 2(12):e190. 2006. [PubMed: 17194218]
28. Pirinen, Matti, Donnelly, Peter, Spencer, Chris CA., et al. E†cient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics. 2013; 7(1):369–390. 2013.
29. Price, Alkes L., Patterson, Nick J., Plenge, Robert M., Weinblatt, Michael E., Shadick, Nancy A., Reich, David. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38(8):904–909. 2006. [PubMed: 16862161]

30. Pritchard, Jonathan K., Donnelly, Peter. Case–control studies of association in structured or admixed populations. Theoretical population biology. 2001; 60(3):227–237. 2001. [PubMed: 11855957]

31. Rakitsch, Barbara, Lippert, Christoph, Stegle, Oliver, Borgwardt, Karsten. A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics. 2013; 29(2):206–214. 2013. [PubMed: 23175758]

32. Segura, Vincent, Vilhjálmsson, Bjarni J., Platt, Alexander, Korte, Arthur, Seren, Ümit, Long, Quan, Nordborg, Magnus. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics. 2012; 44(7):825–830. 2012. [PubMed: 22706313]

33. Tibshirani, Robert. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267–288. 1996.

34. Valdar, William, Solberg, Leah C., Gauguier, Dominique, Burnett, Stephanie, Klenerman, Paul, Cookson, William O., Taylor, Martin S., Nicholas, J., Rawlins, P., Mott, Richard, Flint, Jonathan. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature genetics. 2006; 38(8):879–887. 2006. [PubMed: 16832355]

35. Wang, Haohan, Lengerich, Benjamin J., Aragam, Bryon, Xing, Eric P. Precision lasso: Accounting for Correlations in High-dimensional Genomic Data. 2017 (2017) submitted.

36. Wang, Haohan, Lengerich, Benjamin J., Kyung Lee, Min, Xing, Eric P. GenAMap on Web: Visual Machine Learning for Next-Generation Genome Wide Association Studies. 2017 (2017) submitted.

37. Wang, Haohan, Yang, Jingkang. Multiple Confounders Correction with Regularized Linear Mixed Effect Models, with Application in Biological Processes. Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. 2016 2016.

38. Wang, Jialei, Fujimaki, Ryohei, Motohashi, Yosuke. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2015. Trading interpretability for accuracy: Oblique treed sparse additive models; p. 1245-1254.

39. Xu, Zhixiang, Huang, Gao, Weinberger, Kilian Q., Zheng, Alice X. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2014. Gradient boosted feature selection; p. 522-531.

40. Zhang, Bin, Gaiteri, Chris, Bodea, Liviu-Gabriel, Wang, Zhi, McElwee, Joshua, Podtelezhnikov, Alexei A., Zhang, Chunsheng, Xie, Tao, Tran, Linh, Dobrin, Radu, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimerfis disease. Cell. 2013; 153(3):707–720. 2013. [PubMed: 23622250]

41. Zhang, Cun-Hui. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics. 2010; 2010:894–942.

42. Zhang, Cun-Hui, Zhang, Tong. A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems. Statist Sci. 2012; 27(4):576–593. 2012.

43. Zhou, Jiayu, Lu, Zhaosong, Sun, Jimeng, Yuan, Lei, Wang, Fei, Ye, Jieping. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2013. Feafiner: biomarker identification from medical data through feature generalization and selection; p. 1034-1042.

44. Zhou, Xiang, Stephens, Matthew. Efficient Algorithms for Multivariate Linear Mixed Models in Genome-wide Association Studies. arXiv preprint arXiv:1305.4366. 2013 2013.

45. Zou, Hui, Hastie, Trevor, Tibshirani, Robert. Sparse principal component analysis. Journal of computational and graphical statistics. 2006; 15(2):265–286. 2006.

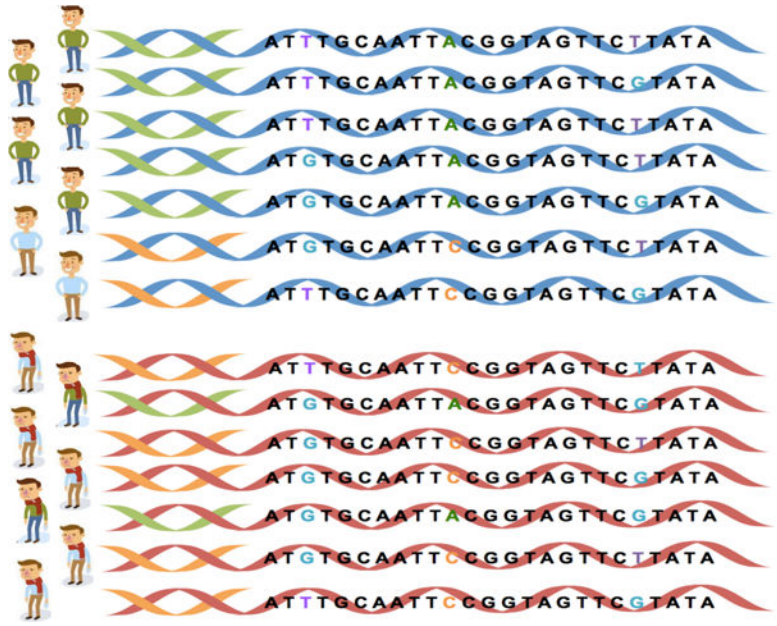

Figure 1.

An illustration of population as a confounding factor in genetic association studies. If the top row comes from one population (e.g. a hospital) and the bottom row comes from a second population (e.g. a different hospital), then population is a confounding factor between disease status and the genetic variables.

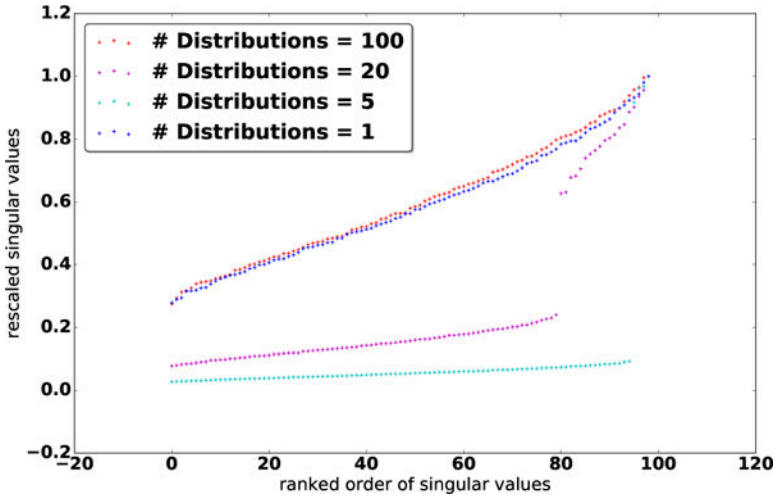

Figure 2.
Distributions of singular values of K for different number of distributions the data originate.

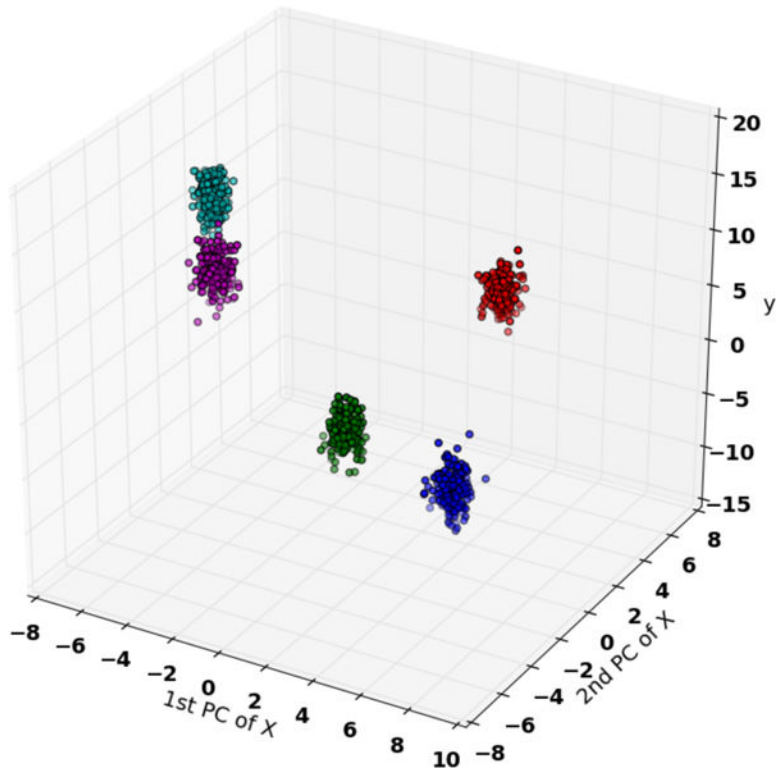

Figure 3.
Synthetic Data: 1000 data samples from five different distributions. Response variables are not only influenced by explanatory variables, but also the distribution it originate from.

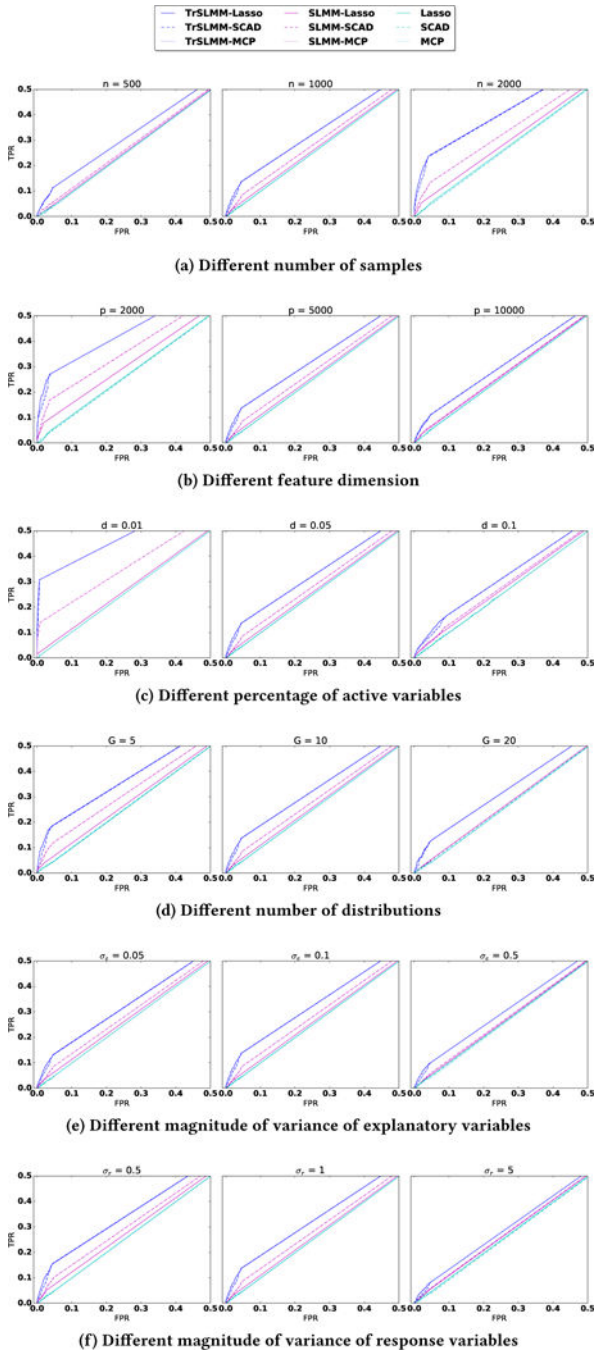

Figure 4.

ROC curves for the variable selection experiment. We have zoomed-in to focus on the region of most interest. For each configuration, the reported curve is drawn over five random seeds.

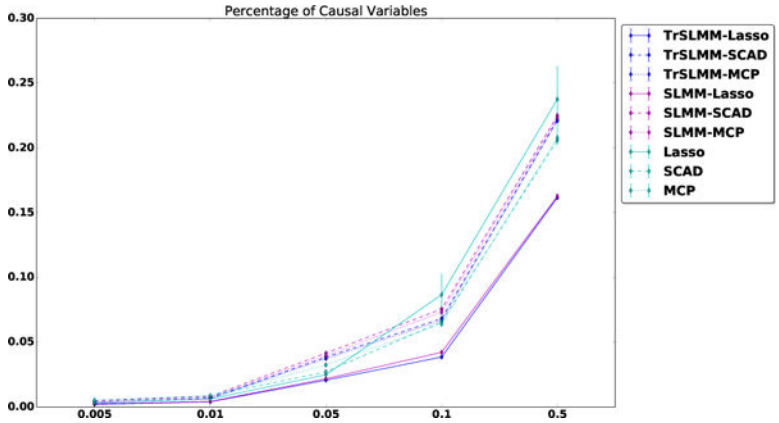

Figure 5.

Mean squared error and its standard error with the prediction of true β

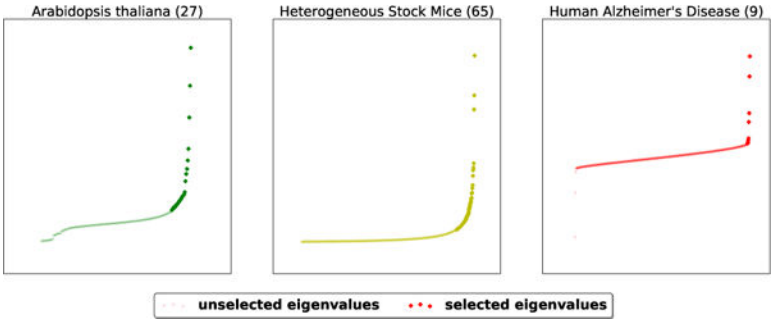

Figure 6.
The selected eigenvalues to consider as the sources of confounding factors

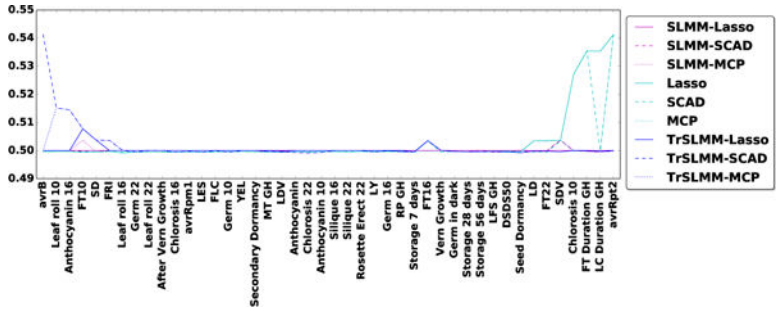

Figure 7.
Area under ROC curve for the 44 traits of Arabidopsis thaliana.

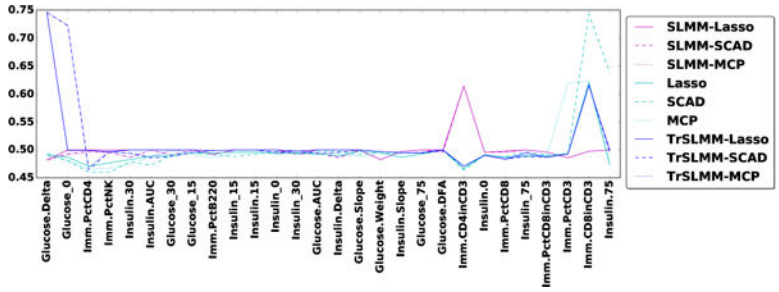

Figure 8.
Area under ROC curve for the 28 traits of Arabidopsis thaliana.

Table 1

Discovered Genetic Variable with TrSLMM-MCP

Rank	SNP	Rank	SNP	Rank	SNP
1	rs10027921	34	rs12506805	67	rs232777
2	rs12641981	35	rs2858166	68	rs4827815
3	rs30882	36	rs4330657	69	rs1518962
4	rs2075642	37	rs3021274	70	rs980354
5	rs12743345	38	rs9632774	71	rs2748423
6	rs12734277	39	rs3788756	72	rs5910006
7	rs388192	40	rs964879	73	rs5927478
8	rs10512516	41	rs7965805	74	rs9290786
9	rs4076941	42	rs2022462	75	rs4828054
10	rs684240	43	rs5906991	76	rs1573980
11	rs4898198	44	rs963658	77	rs4830576
12	rs874404	45	rs1361696	78	rs1921696
13	rs16844380	46	rs5972815	79	rs1921703
14	rs12563627	47	rs12006616	80	rs5927408
15	rs462841	48	rs9377361	81	rs4828057
16	rs12131475	49	rs6950505	82	rs12386970
17	rs1444698	50	rs5920707	83	rs595339
18	rs4243527	51	rs4263905	84	rs8008865
19	rs5907636	52	rs16985176	85	rs8019291
20	rs596997	53	rs5908515	86	rs2816748
21	rs11485173	54	rs6418738	87	rs10838134
22	rs1551055	55	rs16885694	88	rs2414299
23	rs584478	56	rs3788727	89	rs1353821
24	rs9938976	57	rs4824796	90	rs1794648
25	rs5978841	58	rs12006660	91	rs4559363
26	rs6446700	59	rs3747393	92	rs12859216
27	rs9384111	60	rs596720	93	rs1557087
28	rs4421632	61	rs5934013	94	rs13122922

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2018 April 06.

Rank	SNP	Rank	SNP	Rank	SNP
29	rs754865	62	rs6431428	95	rs1186809
30	rs5951621	63	rs2822890	96	rs2225160
31	rs6616026	64	rs4518745	97	rs6802179
32	rs4827217	65	rs4868120	98	rs7018499
33	rs7984044	66	rs7938527	99	rs1009123