UNIFIED SPARSE REGRESSION MODELS FOR SEQUENCE VARIANTS ASSOCIATION ANALYSIS

AN ABSTRACT

SUBMITTED ON THE TWENTY FIFTH DAY OF APRIL, 2015

TO THE DEPARTMENT OF BIOMEDICAL ENGINEERING

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF SCIENCE AND ENGINEERING

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY

Shaolong Cao, B.S.

APPROVED:

Yu-Ping Wang, Ph.D. Director

J. Quincy Brown, Ph.D.

Huaizhen Qin, Ph.D.

Jian Li, Ph.D.

ABSTRACT

Joint adjustment of cryptic relatedness and population structure is necessary to reduce bias in DNA sequence analysis; however, existent sparse regression methods model these two confounders separately. Incorporating prior biological information has great potential to enhance statistical power but such information is often overlooked in many existent sparse regression models. We developed a unified sparse regression (USR) to incorporate prior information and jointly adjust for cryptic relatedness, population structure and other environmental covariates. Our USR models cryptic relatedness as a random effect and population structure as fixed effect and utilize the weighted penalties to incorporate prior knowledge. As demonstrated by extensive simulations, our USR algorithm can discover more true causal variants while maintain a lower false discovery rate than do several commonly used feature selection methods. It can detect rare and common variants with almost equal efficiency.

After further investigation and assessing the oracle property of the USR method, we propose a unified test (uFineMap) for accurately localizing causal loci and a unified test (uHDSet) for identifying high-dimensional sparse associations in deep sequencing genomic data of multi-ethnic individuals. These novel tests are based on scaled sparse linear mixed regressions with L_p ($0<p<1$) norm regularization. Under extensive simulated scenarios, the proposed tests appropriately controlled Type I error rate and appeared more powerful than several existing prominent methods (famSKAT and Gemma).

In addition, we incorporate the idea of Generalized Linear Mixed Models (GLMMs) to further extend the USR model for non-Gaussian phenotype data. The generalized USR method include structure regularization (i.e., group L_1 norm and sparse group L_1 norm) as

well. The algorithm is applicable to a wide range of genetic data association analyses, which can incorporate the effect of a group of SNPs or genes in an integrative way. It can be used as variable screening method to reduce the number of variables, under a wide range of high-dimensional data with complex group structure.

UNIFIED SPARSE REGRESSION MODELS FOR SEQUENCE VARIANTS ASSOCIATION ANALYSIS

A DISSERTATION

SUBMITTED ON THE TWENTY FIFTH DAY OF APRIL, 2015

TO THE DEPARTMENT OF BIOMEDICAL ENGINEERING

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF SCIENCE AND ENGINEERING

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY

Shaolong Cao, B.S.

APPROVED:

Yu-Ping Wang, Ph.D. Director

J. Quincy Brown, Ph.D.

Huaizhen Qin, Ph.D.

Jian Li, Ph.D.

ACKNOWLEDGEMENT

I would like to express my special appreciation and thanks to my advisor Dr. Yu-Ping Wang, who has been a tremendous mentor for me. I would like to thank him for encouraging me and helping me grow as a research scientist. His advice on both research as well as on my career developing have been priceless. I would also like to thank my collaborator Dr. Huaizhen Qin, for his precious guidance and enthusiastic teaching whenever I have a problem. His strong analytical background in biostatistics helped me step into the fields quickly. I also want to thank my committee members, Dr. J. Quincy Brown and Dr. Jian Li for serving as my committee members and great suggestions. In addition, I want to thank them for making my defense be an enjoyable moment, and for the brilliant comments and suggestions. Last but not least, I also appreciate the help from Dr. Hong-Wen Deng, for his insightful suggestions based on his broad knowledge in bioinformatics and genetics. His encouragement on interdisciplinary collaboration with other people in the center for Bioinformatics and Genomics inspired me a lot.

I sincerely appreciate the help from all the alumnus and current members in our Multiscale Bioimaging and Bioinformatics Laboratory (MBB). I want to thank Alexej Gossmann, Dr.Hongbao Cao, Dr.Junbo Duan, Dr. Dongdong Lin, Dr. Jingyao Li, Dr. Chen Qiao, Dr. Jian Fang, Dr. Su-ping Deng, Wenxing Hu, Dr. Pascal Zille, Dr.Jigang Zhang, Dr. Lan-Juan Zhao, Hao He, Chao Xu, Weiwei Ouyang, Xiaoying Fu and Ruifeng Wang for their very useful advices and support in my research. I also want to thank other lab members, Dr. Keith Dillon, Dr. Md Ashad Alam, Yutong Bai, Junqi Wang, Aiying Zhang, Min Wang, Dr. Jie Wu, Dr. Gang Li and Zheng Zhao, for making the lab be such an enjoyable and comfortable place to work. In addition, I want to thanks to those people who supported me at Tulane University, including faculties and staff in both departments of Biomedical Engineering, and Bioinformatics and Biostatistics.

A special thanks to my family. Words cannot express how grateful I am to my wife, Jie Dai, my mother, Xiaoqin Zhao, and my father, Yujun Cao for all of the sacrifices that you've made on my behalf. I undoubtedly could not have done this without your unconditional love and support. I would also like to thank all of other family members and my friends who supported me in writing, and incented me to strive towards my goal.

This work is partially supported by NIH (R01 GM109068, R01 MH104680, R01 MH107354) and NSF (#1539067). The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. Funding support for the Framingham Bone Mineral Density Hip and Spine dataset was provided by NIH grants R01 AR/AG 41398.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Genome wide association study (GWAS)

Genome-wide association studies (GWAS) is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases. GWAS become a relatively common way for scientists to identify genes involved in human disease. The first successful GWAS was published in 2005. It investigated patients with age-related macular degeneration and found two SNPs with significantly altered allele frequency compared to healthy controls (Klein, Zeiss et al. 2005). Next-generation sequencing (NGS) technologies provide great potential for identifying both rare and common sequence variants. A number of GWAS have been developed for identifying marker sets that harbor functional genetic variants.

An illustration of a Manhattan plot (Figure 1.1) depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level (Ikram, Xueling et al. 2010).

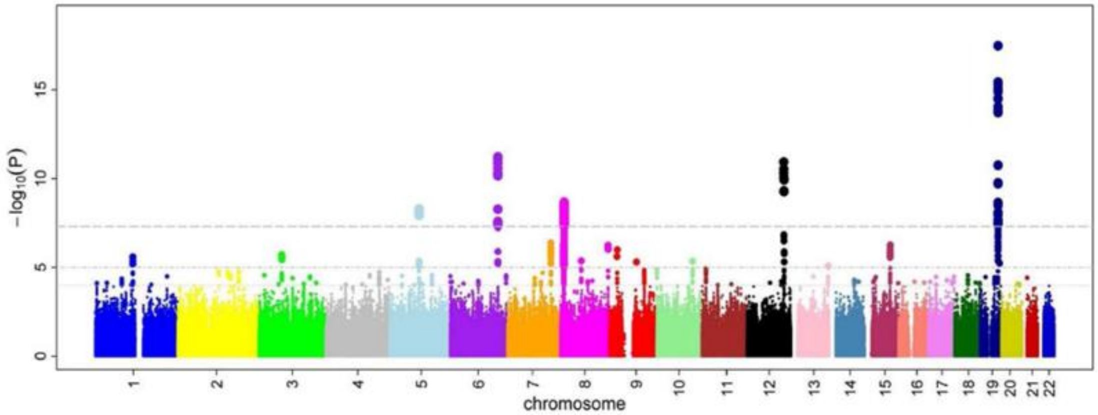

Figure 1.1. Manhattan plot from a GWAS study investigating microcirculation (Ikram, Xueling et al. 2010)

GWAS also have several issues and limitations. Lack of well-defined case and control groups, insufficient sample size, control for multiple testing and control for population stratification are common problems (Pearson and Manolio 2008).

Single marker association tests bear poor statistical power to identify associated rare variants due to their very low frequencies. Reginal or gene set based association tests yield better power. Generalized linear model provides an effective approach to identify variant sets associated with different type of phenotypes while adjust for covariates of unrelated individuals (Yi, Liu et al. 2011, Lee, Wu et al. 2012). However, the assumption of independence between individuals is frequently violated in sequence association studies. In the presence of complex pedigree structure or/and cryptic relatedness, it is more challenging to correct for population structure (Price, Zaitlen et al. 2010), especially for rare variants detection (Mathieson and McVean 2012). Furthermore, some prominent tests (e.g., SKAT family tests and Gemma) require that the number of markers in a testing set is much smaller than the sample size. In a typical population deep sequencing study, it is quite common that interested genomic regions or even the whole genome will have genotypic data of a larger number of marker loci (close to or even larger than sample size, which we call it high dimensional set or HDS), but the functional genetic variants are very sparse among all the variants under test.

1.2 Mixed models

A mixed model is a statistical model containing both fixed effects and random effects. It is particularly useful in settings where repeated measurements are made on the same statistical units (longitudinal study), or where measurements are made on clusters of related statistical units. In another word, the random effect is designed to resolve the non-independence by assuming a different "baseline" value for each subject.

1.2.1 Linear mixed model (LMM)

In most linear model, major interest is on average and variation about the averages. For example, linear regression, T-tests and ANOVA test. The variant of linear mixed model is where parameters in an linear model are treated not as constant but as random variables. Consider the simple model $\mu_{ij} = \mu + \alpha_i + \beta_j$, where α_i is the random effect and β_j represents fixed effect, μ_{ij} is then a mixture of random and fixed term. The model is called linear mixed model (LMM).

A number of sequence association tests have been developed for identifying marker sets that harbor functional genetic variants. Most of them, however, do not jointly model cryptic relatedness, population structure and other covariates. These confounders, if not appropriately adjusted for, may inflate false positive rates or deflate false negative rates.

With the growing demand of analyzing next generation sequencing data of multi-ethnic individuals, linear mixed models are emerging as a method of choice for conducting genetic association studies in humans and other organisms. The advantages of the mixed-model association methods include the prevention of false positive associations due to population or relatedness structure and an increase in power obtained through the application of a correction that is specific to this structure (Yang, Zaitlen et al. 2014). There is a variety of different linear mixed model methods/software packages have been developed for the application of GWA studies, such as EMMAX (Kang, Sul et al. 2010), GenABEL (Aulchenko, Ripke et al. 2007), FaST-LMM (Lippert, Listgarten et al. 2011), Mendel (Lange, Papp et al. 2013), famSKAT (Chen, Meigs et al. 2013) and GEMMA (Zhou and Stephens 2012). These methods differ in details of methodology implemented and various user-chosen options such as the method and number of SNPs used to estimate the kinship (relatedness) matrix (Eu-Ahsunthornwattana, Miller et al. 2014). Within the framework of linear mixed models, famSKAT and GEMMA appeared as two powerful sequence association tests for identifying small marker sets that harbor dense functional genetic variants.

1.2.2 Generalized linear mixed model (GLMM)

The past few decades have seen LM and LMM extended to generalized linear model (GLM) and generalized linear mixed model (GLMM). The essence of this generalization is two-fold: first, the data are not necessarily assumed to be Gaussian distributed; second, that the mean is not necessarily taken as a linear combination of parameters but some link function of mean is. If all the parameters are considered as fixed constants the model is a GLM; if some are treated as random it is a GLMM (McCulloch and Neuhaus 2001).

Recently, some promising association methods are proposed to handle the non-Gaussian phenotype by using GLMM. Lea presented a binomial mixed model and an efficient, sampling-based algorithm (MACAU: Mixed model association for count data via data augmentation) for approximate parameter estimation and p-value computation (Lea, Tung et al. 2015). This framework allows users to simultaneously account for both the over-dispersed, count-based sequencing data, as well as genetic relatedness among individuals. Another savvy GLMM method is lme4 (Bates 2014) which implement several different types of mixed-effects models, including linear mixed models, generalized linear mixed models and nonlinear mixed models.

1.3 Sparse regression

Sparse regression is highly related to sparse representation or sparse approximation in compressive sensing problem. A general goal of sparse regression is to reconstruct a signal or regression coefficients of sampling measurements. Sparse regression is especially powerful for solving underdetermined system and prevent overfitting.

In general, sparse regression is realized by the regularized regression, which refers to the regularized optimization problem $\mathbf{x} = \arg\min_{\mathbf{x}}\{f(\mathbf{x},\mathbf{y}) + P(\mathbf{x})\}$. $f(\mathbf{x},\mathbf{y})$ is the cost function of prediction **x** given data **y**, and $P(\mathbf{x})$ is typically a penalty on the complexity of **x**. Regularization term can be used to learn simpler models, induce models to be sparse, introduce group structure into the learning problem.

1.3.1 Lasso

Lasso (least absolute shrinkage and selection operator), introduced by Robert Tibshirani (Tibshirani 1996), is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

In the linear regression setting, suppose that we are given N samples $\{(\mathbf{x}_i, y_i)\}$, where each $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{im})^T$ is a m-dimensional predictor vector, and $y_i \in R$ is the response variable. It is often write the Lasso problem in the Lagrangian form for some $\lambda > 0$.

$$\min_{\boldsymbol{\beta} \in R^m}\{\frac{1}{2N} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1\}$$

Lasso can be interpreted as linear regression for which the coefficients have Laplace prior distributions. The Laplace distribution is sharply peaked at zero (its first derivative is discontinuous) and it concentrates its probability mass closer to zero than does the normal distribution. This provides an alternative explanation of why lasso tends to set some coefficients to zero. The counter plot of Figure 1.2 demonstrates how the L1 norm penalty induce sparsity. Basically, sparsity is induced by sharp edges lying on the axis of an isosurface.

The lasso problem is a quadratic program with a convex constraint. There are many quadratic program methods for solving the lasso. However, a simple and effective computational algorithm is coordinate descent while applying soft-thresholding for each iteration (Hastie, Tibshirani et al. 2015).

The tuning parameter $\lambda > 0$ controls the complexity of the model. Smaller value of λ induce more parameters and allow the model to have a better fitting to the training data. On the contrary, larger λ restrict the parameters more, leading to sparser, more interpretable models that fit the data less closely. An example of solution path of Lasso with respect to different λ is shown in Figure 1.3.

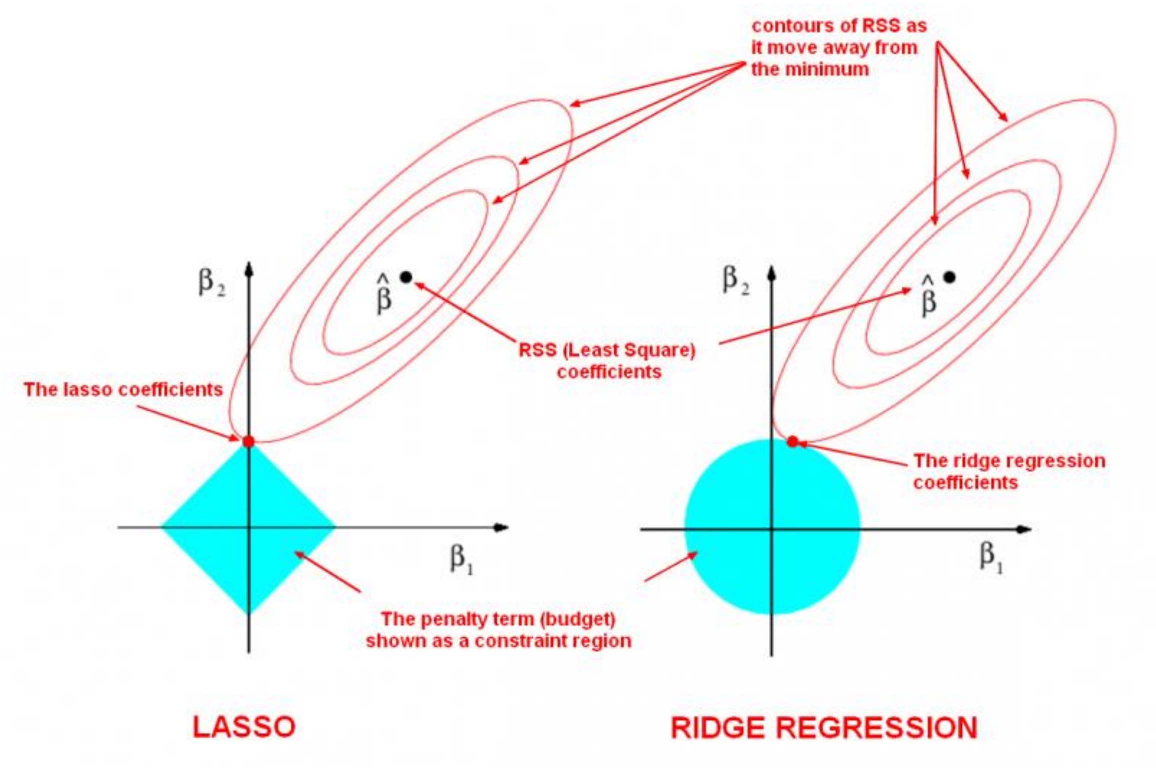

Figure 1.2. Two-dimension contour plot of lasso (left) and ridge regression (right).The solid blue areas are constraint region. The $\hat{\beta}$ depicts the unconstrained least square estimator

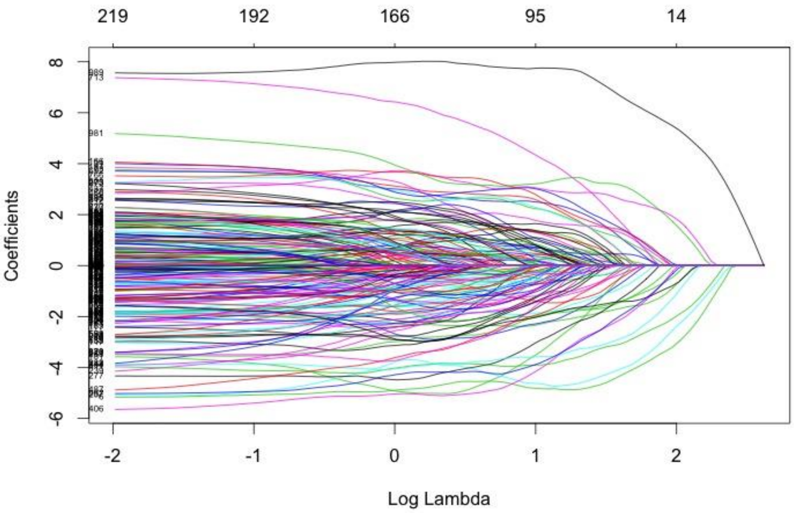

Figure 1.3. Solution path for a Lasso problem

1.3.2 Elastic net

In the fitting of regression models, the elastic net is a regularized regression method that linearly combines the L_1 and L_2 norm penalties of the lasso and ridge methods. Elastic-net is first introduced by (Zou and Hastie 2005) to address several shortcoming of Lasso. When $p > n$ (the number of predictors is greater than the sample size) Lasso can select only n predictors (even when more are associated with the outcome) and it tends to select only one predictor from any set of highly correlated covariates. Additionally, even when $n > p$, if the predictors are strongly correlated, ridge regression tends to perform better. Again, the elastic net can be written by Lagrangian form

$$\min_{\boldsymbol{\beta} \in R^m} \{ \frac{1}{2N} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda_2 \| \boldsymbol{\beta} \|_2^2 + \lambda_1 \| \boldsymbol{\beta} \|_1 \}$$

Elastic net is also a quadratic program with convex constraint, so it can be solved by the similar algorithm as the lasso. It also has been proven that the Elastic net can be reduced to support vector machine (Zhou, Chen et al. 2014).

1.3.3 Group sparsity

Group lasso is a natural extension of the lasso. Yuan and Lin introduced the group lasso in order to allow predefined groups of covariates to be selected into or out of a model together, so that all the members of a particular group are either included or not included (Yuan and Lin 2006). Groups of features can be regularized by a sparsity constraint, which can be useful for expressing certain prior knowledge into an optimization problem.

Another setting in which grouping is natural is in biological studies. Since SNPs and proteins often lie in known gene regions, an investigator may be more interested in which genes are related to an outcome than whether particular individual SNPs are. The objective function for the group lasso is a natural generalization of the standard lasso objective

$$\min_{\boldsymbol{\beta}\in R^m}\{\frac{1}{2N}\|\mathbf{y}-\sum_{j=1}^{J}\mathbf{X}_j\boldsymbol{\beta}_j\|_2^2+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2\}$$

Group lasso is usually solved by block-wise coordinate descent algorithm (Yuan and Lin 2006). The convexity ensure the algorithm can be converged to the global minimal point. It is possible to extend the group lasso to the so-called sparse group lasso (Simon, Friedman et al. 2013), which can select individual covariates within a group, by adding an additional L_1 norm penalty to each group subspace. The objective function of sparse group lasso is shown below.

$$\min_{\boldsymbol{\beta}\in R^m}\{\frac{1}{2N}\|\mathbf{y}-\sum_{j=1}^{J}\mathbf{X}_j\boldsymbol{\beta}_j\|_2^2+\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2+\lambda_2\|\boldsymbol{\beta}\|_1\}$$

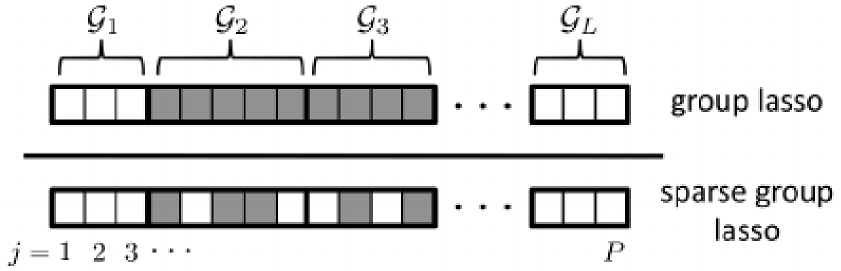

Figure 1.4. An example of selected predictors via group lasso and sparse group lasso regularization term.

1.3.4 Application in genomic data analysis

In recent years, sparse regression models are widely used in genomic data. Given the large number of genetic variants from NGS data and only limited sample size, sparse regression model appeals more powerful than traditional regression based method in term of variants or genes fine mapping. Accurately pinpointing specific causal variants is necessary for elucidating genetic architecture of a complex disease. Sparse representation models were established to select a promising sparse set from a large number of variants(Wu, Chen et al. 2009, Zhou, Sehl et al. 2010) (Wu, Chen et al. 2009, Zhou, Sehl et al. 2010), e.g., those within a gene or a pathway. Such models allow the size of a testing set (gene or pathway) exceed the number of study participants (Fan and Li 2001, Zou and Hastie 2005) by the use of regularization terms (e.g., L_0 norm, L_1 norm, L_2 norm).

Lasso penalized logistic regression was first implemented in case–control disease gene mapping (Wu, Chen et al. 2009). The elastic net regularized regression along with cross-validation to find the optimal tuning parameter was investigated in GWAS data (Waldmann, Mészáros et al. 2013). (Zhou, Sehl et al. 2010) first applied group lasso in GWAS on breast cancer data. (Larson and Schaid 2014) and (Ayers and Cordell 2013) further extend the structure sparse regularized methods to rare variants analysis. The proposed method analyzes all genes at once, allowing grouping of all (rare and common) variants within a gene, along with subgrouping of the rare variants.

The reminder of the dissertation proposed novel sparse regression methods based on linear mixed model and generalized linear mixed model with L_p ($0<p<1$) norm regularization and group sparse regularization. Although the L_0 norm penalty yields sparsest solution, its discontinuity makes the problem to be NP-hard (Natarajan 1995), which is nearly infeasible for the regression model with a large number of predictors. The L_1 norm penalty or Lasso is a well-developed and computationally feasible method, with the relaxation of L_0 norm penalty. On one hand, if a particular restricted isometric property (RIP) holds, the solution of lasso and L_0 norm penalty are identical (Candes and Tao 2005); on the other hand, solving L_1 norm is a convex optimization problem which is feasible for large scale genetic variants. Elastic-net (Zou and Hastie 2005) is a mixture penalty derived from L_1 norm and L_2 norm. Although it is more robust to Gaussian noise than Lasso, the additional turning parameter requires more computational effect to estimate a global optimal solution. Recently, L_p norm ($0 < p < 1$), as an alternative relaxation, has aroused more interests (Xu, Chang et al. 2012), which yields more sparse solutions than does the Lasso. Despite these merits, existent sparse representation algorithms still suffer the limitations of aforesaid set (e.g., gene, pathway) based association methods.

Figure 1.5 and Figure 1.6 show the unit balls and comparison of different L_p norms. Sparse solution requires a L_p ($0<p\leq1$) norm.

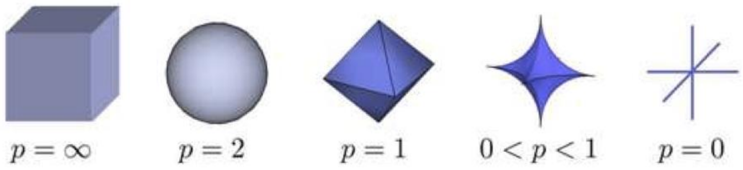

Figure 1.5 Unit ball for different L_p norms in 3-dimension

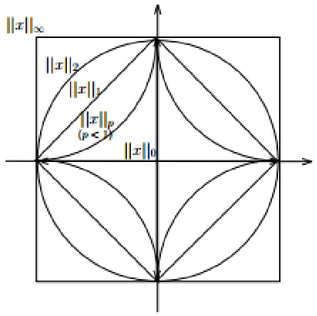

Figure 1.6 Unit ball for different L_p norms in 2-dimension

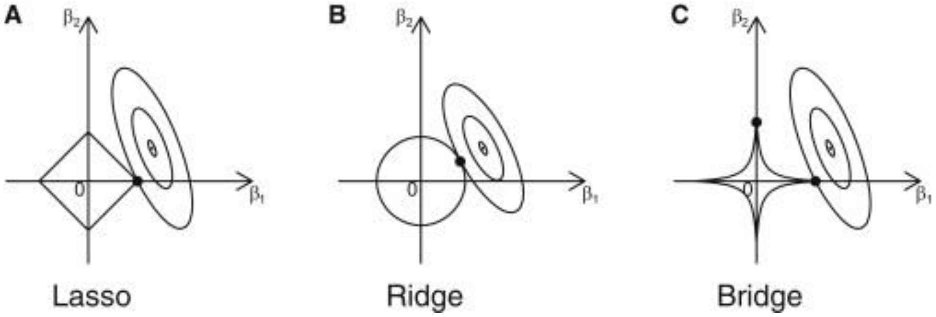

Figure 1.7 two-dimension contour plot of lasso, ridge and bridge regression (L_p ($0<p\leq1$) norm)

CHAPTER 2 THE UNIFIED SPARSE REGRESSION (USR) MODEL

2.1 Introduction

Complex diseases and traits are likely to be influenced by both common and rare genetic variants. The whole genome sequencing has provided a powerful tool for the study of complex diseases and traits, because more rare variants can be detected from the sequencing data with higher resolutions. In other words, the number of rare variants detected by sequencing data is much larger than that of common variants detected in GWAS studies. In addition, the low frequency of rare variants makes association based testing extremely difficult, i.e., having the low statistical power to detect each single rare variant. For this reason, current methods use two major approaches. First, one collapses or combines the genotype data in a specific region, then claim whether the whole region is associated with traits or not. Although with a considerably power gain, these kinds of methods lack the ability to pinpoint the causal variants at SNP level. The second approach is to detect rare and common variants separately or set a threshold based on minor allele frequency. However, such a test approach is in favor of common variants, which usually lose power for detecting rare variants, and vise versa. We aim to provide a method that can consider both rare and common variants equally while maintain the accuracy of single causal variants detection.

A generalized linear model (GLM) provides a popular way to take both common and rare variants as well as environmental factors into consideration. Nevertheless, the classical GLM has limitations. First, GLM tends to detect too many variants corresponding to non-zero coefficients in the regression model, resulting in too many degrees of freedom and the reduction of statistical power. It is hard to pinpoint which variants are the main causal variants. Second, GLM assumes that individuals in the samples are unrelated, i.e., independently identical distributed (i.i.d), which is not the case in practice. A model that can adjust complex pedigree structures and handle high dimensional variants data is needed. To this end, we propose a novel sparse regression model to overcome the limitations of existing methods.

Classical sparse regression models always assume that the individuals in data are i.i.d, so that the quadratic loss function is an unbiased estimator of likelihood and acts as the loss function. However, this is not the case in real world. According to the pedigree's impact on continuous phenotypes (Thompson and Shaw 1990), we propose to use a modified Kinship matrix to adjust the correlation between pedigrees and develop a new quadratic loss function. In short, our model can deal with the data of complex relationship and pedigree structures, which are common in real case.

In addition, we want to utilize as much prior information as possible. A natural way to incorporate prior knowledge is to add a weight coefficient on each feature. By doing this, we encourage the highly suspected variants while discourage lower risk variants into the model. This design gives a flexible framework in variants selection.

In this dissertation, we propose a sparse regression model with adjustment of pedigree structure and with weighted sparse penalty terms including Elastic-net, and L_p norm ($0<p<1$) in order to detect the association of genotype with phenotype data. The phenotype data can be either continuous or binary. We solve the $L_{1/2}$ norm problem by the half threshold algorithm (Xu, Chang et al. 2012), and solve the L_p norm regularization model by a smoothing method (Chen, Xu et al. 2010). Our modified elastic-net regularization model is solved by a coordinate decent algorithm (Friedman, Hastie et al. 2007, Friedman, Hastie et al. 2010), which is faster than many existing algorithms especially when the data matrix is large. Based on the solution path for different regularization parameter λ , we use the Akaike Information Criteria (AIC) to choose an optimal penalty parameter λ . Then we use the stability selection method to get the appropriate sparse regression coefficients. To evaluate our methods, we compare our method with the single marker test (χ^2 test), Elastic-net, Orthogonal Matching Pursuit (OMP) and FOcal Underdetermined System Solver (FOCUSS) (Rao and Kreutz-Delgado 1999). Furthermore, we extend our family adjustment and weighted model to the one with Elastic-net penalty.

Our proposed approach has the following advantages: (i) The model can adjust pedigree structures; (ii) The L_p norm regularization model can yield higher true positive rate while lower false discovery rate than other methods; (iii) The weighted regularization term provides a flexible way to incorporate prior knowledge; (iv) Our model can be easily extended to accommodate environmental covariates.

The reminder of the dissertation is organized as follows. In the method section, we present our family adjusted sparse regression model with weighted regularization in detail. In the result section, we evaluate the performance of our model on both simulation and real data. In the conclusion and discussion section, we discuss both the advantages and disadvantages of our model. Finally, we give a perspective on our future work.

2.2 The USR (Unified sparse regression) method

2.2.1 Notations

Let n denote the total number of subjects, and m denote the number of independent variables. Let $\mathbf{Y} = (y_1, y_2, ..., y_n)^T$ contain the trait values of the n subjects. We write

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^T$, represents genotype data for subject *i*; $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_L)$, where $\mathbf{w}_i = (w_{i1}, w_{i2}, ..., w_{in})^T$ represents fixed-effect confounders, e.g., population structure surrogates, age, and gender; and $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)^T$, where $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{in})^T$ represents random-effect, e.g., pedigree structure.

2.2.2 Joint adjustment of confounders

For data with a known pedigree structure, we consider the following linear mixed-effect model:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}_0 \qquad (2.1)$$

where $\boldsymbol{\varepsilon}_0 \sim N(\mathbf{0}, \mathbf{I}_n)$, $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Phi})$, $\boldsymbol{\Phi} = (\Phi_{ij})$ is the kinship matrix or IBD (Identity-by-Descent) matrix, Φ_{ij} equals to twice of the kinship coefficient between subject *i* and *j*; $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_L)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_m)^T$ are vectors of corresponding regression coefficients.

A classical model was proposed by (Thompson and Shaw 1990) to summarize the random effect due to pedigree structure and random noise $\boldsymbol{\varepsilon}_0$ into a new error term $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \sigma_{\Phi}^2 \boldsymbol{\Phi} + \sigma_{\varepsilon}^2 \mathbf{I}$

To be explicit:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2.2)$$

For the data of cryptic relatedness, the kinship matrix can be inferred by extent algorithm, e.g., the REAP (Thornton, Tang et al. 2012). The coefficients of variance component σ_{Φ}^{2} and σ_{ε}^{2} can be inferred by

For a given $\mathbf{\Sigma}$, the likelihood can be formulated as:

$$L(\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\mathbf{\Sigma}|}}\exp(-\frac{(\mathbf{Y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})^{T}\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})}{2})$$

The Log-likelihood is:

$$l(\boldsymbol{\alpha},\boldsymbol{\beta})=-\log(L(\boldsymbol{\alpha},\boldsymbol{\beta}))\propto(\mathbf{Y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})^{T}\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})$$

2.2.3 The generic L_p regularization

A general form of regularized regression is given by:

$$(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})=\underset{\alpha,\beta}{\arg\min}\{-\log(likelihood(\boldsymbol{\alpha},\boldsymbol{\beta}))+P_{\lambda}(\boldsymbol{\beta})\}$$

The natural approach towards regularizing the sparsity of the solution of to use the number of non-zero coefficients as a penalty, i.e., L_0 norm. However, it is not computationally tractable. L_p norm is a closer relaxation comparing to L_1 norm. It is well known (Chen, Xu et al. 2010, Xu, Chang et al. 2012) that L_p (0<p<1) norm regularization term can give more sparse solution than L_1 norm based regularization, also known as the famous least absolute shrinkage and selection operator (Lasso). If we define the L_p norm based regularization term as $P_{\lambda}(\boldsymbol{\beta})=\lambda\|\boldsymbol{\beta}\|_{p}^{p}=\lambda\sum_{j=1}^{m}|\beta_{j}|^{p}$, $0<p<1$ then the problem becomes to find the minimizer

$$(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})=\underset{\boldsymbol{\alpha}\in R^{L},\boldsymbol{\beta}\in R^{m}}{\arg\min}\ f(\boldsymbol{\alpha},\boldsymbol{\beta})=\underset{\boldsymbol{\alpha}\in R^{L},\boldsymbol{\beta}\in R^{m}}{\arg\min}\{(\mathbf{Y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})^{T}\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_{p}^{p}\} \tag{2.3}$$

In particular, if the data only contain unrelated subjects, i.e., $\mathbf{\Phi} = \mathbf{I}$, $\mathbf{\Sigma} = (\sigma_{\Phi}^2 + \sigma_{\varepsilon}^2)\mathbf{I}$, eq. (2.3) collapses to the classic least square sparse regression. Similar to other sparse regressions, we define the selected risk variants to be the set of non-zero regression coefficients, i.e., $\{i \mid \beta_i \neq 0\}$.

2.2.4 Incorporating prior information

The regularization term in eq. (2.3) can be modified to incorporate prior knowledge. For this purpose, we introduce a weighted regularization term. To be explicit, the weighted L_p norm regularization is:

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\alpha} \in R^L, \boldsymbol{\beta} \in R^m}{\arg\min} \{(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}) + \lambda \| \boldsymbol{\gamma\beta} \|_p^p\} \tag{2.4}$$

Where $\boldsymbol{\gamma\beta} = (\gamma_1\beta_1, \gamma_2\beta_2, ..., \gamma_m\beta_m)^T$ and $\gamma_j > 0$ represent marker wise weights.

An appropriate choice of weights can improve statistical power. Each weight γ_j is pre-specified, taking the genotypes, covariates and prior knowledge into account. The weight γ_j reflects the relative importance or preference of the *j*th variant. On one hand, we can assign a particular marker with small penalty weight, if we want to include the marker into the sparse representation. On the other hand, a marker with a large weight is more likely to be excluded from the sparse representation.

There are several ways to determine the weights. For example, we can give non-synonymous SNPs or the SNPs in the risk gene lower weights to increase their chances to enter the model. Another way to assign weights is based on minor allele frequency. When analyzing rare and common variants together, we can assign lower weights to rare variants, in order to compensate for their low frequencies. Since in practice we do not know exactly which variants have high risk, the weights should be assigned prudently.

In particular, if all $\gamma_j = 1$, eq. (2.4) collapses to eq. (2.3), which is an unweighted one. The algorithm to solve eq. (2.4) is almost the same as for eq. (2.3). The only difference is to replace β_j by $\gamma_j \beta_j$. For the sake of simplicity, we just present the algorithm for solving eq. (2.3). We assume all variants are equally weighted unless otherwise stated.

2.2.5 Surrogate function of the USR problem

Generally, the L_p ($0<p<1$) norm based regularization (eq. 2.3) is neither convex nor Lipschitz continuous, making the solution computationally difficult and time-consuming. We adopt the basic idea on non-convex and non-continuous optimization (Zhang and Chen 2009) to solve the minimization problem of (2.3). To make the algorithm more stable and faster, we establish a lower bound to further regularize local optimal solution. Another issue with L_p norm regularization is that the iterative algorithm can be easily trapped at a local minimizer. Therefore, the choice of the initial point is crucial for the iterative algorithm. For this reason, we use the solution of $L_{0.5}$ norm regularization as the initial point for the L_p regularization problem. The details are discussed below.

We use a smoothing approximation to the objective function in eq.(2.3) (Chen, Xu et al. 2010)

$$f_\mu(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}) + \lambda \| \psi_\mu(\boldsymbol{\beta}) \|_p^p$$

where $\psi_\mu(\boldsymbol{\beta}) = (s_\mu(\beta_1), s_\mu(\beta_2), ..., s_\mu(\beta_m))^T$ and

$$s_\mu(x) = \begin{cases} |x| & |x| > \mu \\ \dfrac{x^2}{2\mu} + \dfrac{\mu}{2} & |x| \le \mu \end{cases}$$

The smoothing function $f_{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is continuously differentiable and strictly convex on the set of $\{x \mid \max(x) \le \mu\}$. Moreover, $\lim_{\mu \downarrow 0} f_{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta}) = f(\boldsymbol{\alpha},\boldsymbol{\beta})$

All these properties show that this smoothing function is a good approximation to the original one but makes the problem easy to solve.

2.2.6 Lower bound theory

Based on the first and second order necessary condition on the solution to the minimization problem, we derive a lower bound, and a sufficient and necessary condition to narrow the search of non-zero entries and guide the selection of causal variants. In our algorithm, we utilize the lower bound at each step to help refine the local minimizer.

The lower bound theory for the unified sparse model

Denote X_p^* the set of local minimizers of objective formula (2.3)

For any $\boldsymbol{\beta}^* \in X_p^*$ derived from initial point $\boldsymbol{\beta}^0$, the following statements hold

(i) Let $L_i = \max[(\frac{\lambda p(1-p)}{2(\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})_{ii}})^{\frac{1}{2-p}}, (\frac{\lambda p \sqrt{k}}{2\|\mathbf{A}\| \cdot \|\boldsymbol{\Sigma}^{-1}\| \sqrt{f(\boldsymbol{\alpha},\boldsymbol{\beta}^0)}})^{\frac{1}{1-p}}]$

for any $\boldsymbol{\beta}_i^* \in (-L_i, L_i) \Leftrightarrow \boldsymbol{\beta}_i^* = 0$

where $\mathbf{A} := \mathbf{X}_{\Lambda} \in R^{n \times |\Lambda|}$ is a sub-matrix of $\mathbf{X}$, which consists of the jth columns of $\mathbf{X}$, with $j \in \Lambda$, $\Lambda = \text{support}(\boldsymbol{\beta}_i^*) = \{i \mid \boldsymbol{\beta}_i^* \neq 0\}$, $K = \|\boldsymbol{\beta}^*\|_0$

(ii) The smallest eigenvalue of matrix $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}^{-1}$: $\lambda_{\min} \ge 1$;

where $\widetilde{\mathbf{A}} = 2\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A}$ and $\widetilde{\mathbf{B}} = \lambda p(1-p) diag(|\beta_i^*|^{p-2})$, and $\beta_i^* \neq 0$

Proof:

For $\boldsymbol{\beta}^* \in \mathbf{X}_p^*$, $\|\boldsymbol{\beta}^*\|_0 = k$, without loss of generality, we can assume $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, ..., \beta_k^*, 0, ..., 0)^T$

Let $\mathbf{z}^* = (\beta_1^*, \beta_2^*, ..., \beta_k^*)^T$ and $\mathbf{A} \in R^{n \times k}$ be the submatrix of $\mathbf{X}$, whose column is the corresponding index of vector $\mathbf{Z}^*$

Define a function $g: R^k \to R$ by

$$g(\mathbf{z}) := (\mathbf{Y} - \mathbf{W\alpha} - \mathbf{Az})^T \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{W\alpha} - \mathbf{Az}) + \lambda \| \mathbf{z} \|_p^p$$

Intuitively, we have

$$f(\boldsymbol{\beta}^*) = (\mathbf{Y} - \mathbf{W\alpha} - \mathbf{X}\boldsymbol{\beta}^*)^T \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{W\alpha} - \mathbf{X}\boldsymbol{\beta}^*) + \lambda \| \boldsymbol{\beta}^* \|_p^p$$

$$= (\mathbf{Y} - \mathbf{W\alpha} - \mathbf{Az}^*)^T \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{W\alpha} - \mathbf{Az}^*) + \lambda \| \mathbf{z} \|_p^p = g(\mathbf{z}^*)$$

where $\mathbf{z}^*$ is the local minimizer of $g(\mathbf{z})$, i.e., $g(\mathbf{z}^*)$ should satisfy the following second order necessary condition at $\mathbf{z}^*$

$\frac{\partial^2 g(\mathbf{z})}{\partial \mathbf{z}^2} |_{\mathbf{z}=\mathbf{z}^*} = 2\mathbf{A}^T \mathbf{\Sigma}^{-1} \mathbf{A} + \lambda p(p-1) diag(| \mathbf{z}^* |^{p-2})$ should be positive semi-definite.

(i) $2e_i^T \mathbf{A}^T \mathbf{\Sigma}^{-1} \mathbf{A} e_i + \lambda p(p-1) | z_i^* |^{p-2} \geq 0, i = 1,2,..., k$

Note $e_i^T \mathbf{A}^T \mathbf{\Sigma}^{-1} \mathbf{A} e_i = (\mathbf{A}^T \mathbf{\Sigma}^{-1} \mathbf{A})_{ii}$

$$| z_i^* | \geq \left(\frac{\lambda p(1-p)}{2(\mathbf{A}^T \mathbf{\Sigma}^{-1} \mathbf{A})_{ii}}\right)^{\frac{1}{2-p}} = L_1$$

The local minimizer of $g(\mathbf{z}^*) \leq f(\boldsymbol{\beta}^0)$, should satisfy the first order necessary condition at $\mathbf{z}^*$

$$\frac{\partial g(\mathbf{z})}{\partial \mathbf{z}} |_{\mathbf{z}=\mathbf{z}^*} = 2\mathbf{A}^T \mathbf{\Sigma}^{-1} (\mathbf{W\alpha} + \mathbf{Az}^* - \mathbf{Y}) + \lambda p(sign(\mathbf{z}) | \mathbf{z}^* |^{p-1}) = 0$$

Because of $\boldsymbol{\beta}^* \in \mathbf{X}_p^*$,

we have $\lambda p \| | \mathbf{z}^* |^{p-1} \| = 2 \| \mathbf{A}^T \mathbf{\Sigma}^{-1} (\mathbf{W\alpha} + \mathbf{Az}^* - \mathbf{Y}) \|$

$$\begin{aligned}
&\| \mathbf{A}^T \boldsymbol{\Sigma}^{-1} (\mathbf{W}\boldsymbol{\alpha} + \mathbf{A}\mathbf{z}^* - \mathbf{Y}) \|_2^2 \leq \| \mathbf{A}^T \|_2^2 \| \boldsymbol{\Sigma}^{-1} \|_2^2 \| (\mathbf{W}\boldsymbol{\alpha} + \mathbf{A}\mathbf{z}^* - \mathbf{Y}) \|_2^2 \\
&= \| \mathbf{A}^T \|_2^2 \| \boldsymbol{\Sigma}^{-1} \|_2^2 \| (\mathbf{X}\boldsymbol{\beta}^* + \mathbf{W}\boldsymbol{\alpha}^* - \mathbf{Y}) \|_2^2 \\
&\leq \| \mathbf{A}^T \|_2^2 \| \boldsymbol{\Sigma}^{-1} \|_2^2 \| (\mathbf{X}\boldsymbol{\beta}^* + \mathbf{W}\boldsymbol{\alpha}^* - \mathbf{Y}) \|_2^2 + \lambda \| \boldsymbol{\beta}^* \|) \\
&= \| \mathbf{A}^T \|_2^2 \| \boldsymbol{\Sigma}^{-1} \|_2^2 \ f(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \\
&\leq \| \mathbf{A}^T \|_2^2 \| \boldsymbol{\Sigma}^{-1} \|_2^2 \ f(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^0)
\end{aligned}$$

where $\boldsymbol{\alpha}^* \in \arg\min_{\alpha \in R^L} \| \mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} \|^2$

Together with the above formula

$$2 \| \mathbf{A}^T \|_2 \| \boldsymbol{\Sigma}^{-1} \|_2 \sqrt{f(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^0)} \geq \lambda p \| | \mathbf{z}^* |^{p-1} \| \geq \lambda p \sqrt{k} \min_{1 \leq i \leq k} | z_i^* |^{p-1}$$

Finally, we have

$$\min_{1 \leq i \leq k} | z_i^* | \geq \left(\frac{\lambda p \sqrt{k}}{2 \| \mathbf{A}^T \| \| \| \boldsymbol{\Sigma}^{-1} \| \sqrt{f(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^0)}} \right)^{\frac{1}{1=p}} = L_2$$

So the absolute value of all the nonzero entries of the solution should be no less than L_1 and L_2

$$\beta_i^* = \max\left[\left(\frac{\lambda p (1-p)}{2(\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})_{ii}} \right)^{\frac{1}{2-p}}, \left(\frac{\lambda p \sqrt{k}}{2 \| \mathbf{A} \| \cdot \| \boldsymbol{\Sigma}^{-1} \| \sqrt{f(\boldsymbol{\alpha}, \boldsymbol{\beta}^0)}} \right)^{\frac{1}{1-p}} \right]$$

(ii) Define $\widetilde{\mathbf{A}} = 2\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A}$ and $\widetilde{\mathbf{B}} = \lambda p (1-p) diag(| \beta_i^* |^{p-2})$

We have $\widetilde{\mathbf{A}} - \widetilde{\mathbf{B}} \geq 0$, where $\widetilde{\mathbf{A}} > 0, \widetilde{\mathbf{B}} > 0$

Since $\widetilde{\mathbf{B}} > 0$, there exists a unique non-singular matrix $\widetilde{\mathbf{B}}^{\frac{1}{2}}$ such that $\widetilde{\mathbf{B}} = \widetilde{\mathbf{B}}^{\frac{1}{2}} \widetilde{\mathbf{B}}^{\frac{1}{2}}$

Furthermore, we have

$$\widetilde{\mathbf{A}} - \widetilde{\mathbf{B}} = \widetilde{\mathbf{B}}^{\frac{1}{2}} \widetilde{\mathbf{B}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{B}}^{-\frac{1}{2}} \widetilde{\mathbf{B}}^{\frac{1}{2}} - \widetilde{\mathbf{B}}^{\frac{1}{2}} \widetilde{\mathbf{B}}^{\frac{1}{2}} = \widetilde{\mathbf{B}}^{\frac{1}{2}} (\widetilde{\mathbf{B}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{B}}^{-\frac{1}{2}} - \mathbf{I}) \widetilde{\mathbf{B}}^{\frac{1}{2}} \geq 0$$

namely, for any $\mathbf{x} \in R^k$,

$$\mathbf{x}^T \widetilde{\mathbf{B}}^{\frac{1}{2}} (\widetilde{\mathbf{B}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{B}}^{-\frac{1}{2}} - \mathbf{I}) \widetilde{\mathbf{B}}^{\frac{1}{2}} \mathbf{x} = \mathbf{c}^T (\widetilde{\mathbf{B}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{B}}^{-\frac{1}{2}} - \mathbf{I}) \mathbf{c} \geq 0, \text{ where } \mathbf{c} = \widetilde{\mathbf{B}}^{\frac{1}{2}} \mathbf{x}$$

Obviously, there exists an orthogonal matrix $\mathbf{V}$ such that $\tilde{\mathbf{B}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-\frac{1}{2}} - \mathbf{I} = \mathbf{V}^T(\mathbf{\Lambda} - \mathbf{I})\mathbf{V} \geq 0$,

where $\mathbf{\Lambda}$ is the diagonal matrix consisting of eigenvalues of $\tilde{\mathbf{B}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-\frac{1}{2}}$

So all the eigenvalues of $\tilde{\mathbf{B}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-\frac{1}{2}}$ have to be larger than 1, i.e., the solution of the eigenvalue equation are not smaller than 1. To express explicitly,

$|\tilde{\mathbf{B}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-\frac{1}{2}} - \lambda\mathbf{I}| = 0$, where all eigenvalues $\lambda \geq 1$

Multiplying $|\tilde{\mathbf{B}}^{\frac{1}{2}}|$ at the left side, and then multiplying $|\tilde{\mathbf{B}}^{-\frac{1}{2}}|$ at the right side of the equation, we have

$|\tilde{\mathbf{B}}^{\frac{1}{2}}||\tilde{\mathbf{B}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-\frac{1}{2}} - \lambda\mathbf{I}||\tilde{\mathbf{B}}^{-\frac{1}{2}}| = |\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-1} - \lambda\mathbf{I}| = 0$

Consequently, the smallest eigenvalue of matrix $\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{-1} : \lambda_{\min} \geq 1$ □

The more detailed proof of this theory is described in the appendix of USR tests (Cao, Qin et al. 2015).

2.2.7 USR algorithm

The *L0.5* norm regularization has an analytical threshold operator (Xu, Chang et al. 2012) compared with arbitrary *Lp* (0<*p*<1) norm problem, which can be easily and fast solved. In addition, the *L0.5* regularization always yields more sparse solution than that of using *Lp* when 0.5<*p*<1, and shows no significant difference from the one when 0<*p*<0.5 (XU, GUO et al. 2012). Thus in our algorithm, we first apply *L0.5* thresholding algorithm (Xu, Chang et al. 2012) to obtain the solution of *L0.5* problem and then use the solution of *L0.5* as the initial point to search the minimizer of the *Lp* norm based regularization problem.

The *L0.5* regularization model is given by the following formulation (2.5)

$$(\boldsymbol{\alpha}_{opt}, \boldsymbol{\beta}_{opt}) = \underset{\boldsymbol{\alpha}\in R^L, \boldsymbol{\beta}\in R^m}{\arg\min}\{(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_{1/2}^{1/2}\} \tag{2.5}$$

where $\| \boldsymbol{\beta} \|_{1/2} = (\sum_{j=1}^{m} \sqrt{|\beta_j|})^2$

According to (Xu, Chang et al. 2012), the solution of (2.5) can be obtained by the following thresholding operation

$$\boldsymbol{\beta}^* = R_{\lambda\mu,1/2}(\boldsymbol{\beta}^* + \mu \mathbf{X}^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha}^* - \mathbf{X}\boldsymbol{\beta}^*))$$

where $R_{\lambda\mu,1/2}(t)$ is the half thresholding operator. It is given as follows:

$$R_{\lambda\mu,1/2}((x_1, x_2, ..., x_m)^T) = (f_{\lambda\mu,1/2}(x_1), f_{\lambda\mu,1/2}(x_2), ..., f_{\lambda\mu,1/2}(x_m))^T$$

where $f_{\lambda\mu,1/2}(x) = \begin{cases} \frac{2}{3} x(1 + \cos(\frac{2\pi}{3} - \frac{2}{3}\varphi_{\lambda\mu}(x))) & |x| > \frac{\sqrt[3]{54}}{4}(\lambda\mu)^{\frac{2}{3}} \\ 0 & otherwise \end{cases}$

and $\varphi_{\lambda\mu}(x) = \arccos(\frac{\lambda\mu}{8}(\frac{|x|}{3})^{-\frac{3}{2}})$, $\mu = \| \mathbf{X} \|_2$

We name our algorithm to solve the problem (2.5) as the Hybrid *L0.5*-SCG algorithm, where SCG stands for the smoothing conjugate gradient.

Unified sparse regression algorithm

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…*m*

Step 2: For any given λ, *p*, set iterative index *r*=0, ε =0.0001; Initialize $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$, $\boldsymbol{\beta}^{(0)} = \mathbf{0}$

Step 3: Update $\boldsymbol{\alpha}^{(r+1)} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(r)})$

Update $\boldsymbol{\beta}^{(r+1)} = R_{\lambda\mu,1/2}(\boldsymbol{\beta}^{(r)} + \mu\mathbf{X}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha}^{(r+1)} - \mathbf{X}\boldsymbol{\beta}^{(r)}))$

Step 4: Apply the lower bounds to regularize $\boldsymbol{\beta}^{(r+1)}$ and use the SCG algorithm (Zhang *et al.*, 2009) with the initial point $\boldsymbol{\beta}^{(r+1)}$ to find the minimizer $\boldsymbol{\beta}_p^{(r+1)}$ of objective function (2.3)

Step 5: Calculate $\| \boldsymbol{\beta}_p^{(r+1)} - \boldsymbol{\beta}_p^{(r)} \|_2$

If $\| \boldsymbol{\beta}_p^{(r+1)} - \boldsymbol{\beta}_p^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step 3

Then $\boldsymbol{\beta}_p^{(r+1)}$ is the final solution

2.2.8 Tuning parameter selection

It is well known that the setting of regularization (tuning) parameter λ in eqs. (2.3) and (2.5) controls the tradeoff between data fitting fidelity and the use of prior knowledge. A larger λ results in a more sparse solution and vice versa.

The selection of optimal regularization parameters is a difficult problem. If computing time is not a concern, it is helpful to optimize the objective function over a grid of points and monitor how new predictors enter the model as λ decreases. Another way is to minimize either the Bayesian information criterion (BIC) or AIC as a function of λ. Also, we can use cross-validation to select optimal λ. After the comparison of these methods in our simulations, we choose the AIC as our variable selection criterion. For our model, we have the following form of AIC (Cetin and Erar 2002)

$$AIC = 2k + n(\log((\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})) + 1)$$

The goal is to find an optimal λ so that the AIC value can be minimized. Since λ is the key parameter to determine the sparsity level, it is crucial to understand the relationship between AIC and λ. However, there is no explicit expression of AIC(λ). So we use the discrete search in log-scale to find the optimal λ that yields the smallest AIC value.

However, a major drawback of AIC procedure is that it cannot control false positive rate or family-wise error rate. So we use the idea of stability selection (Meinshausen and Bühlmann 2010) to further control the false positive rate based on the selected λ.

The basic idea about stability selection is to bootstrap the data, and then calculate the frequency of the variables to be selected. The higher frequency of the selected variables implies that they are more important. Hence, we can develop a new rank of importance of each variable (i.e., variants), and then a frequency threshold is applied to select final risk variants. An advantage of the stability selection over AIC selection is that the expected number of falsely selected variables or false positive rate can be asymptotically controlled. The detailed procedure of hybrid AIC and stability selection is described below.

Let $I = \{1,2,...,n\}$ index the entire sample, and $S \subseteq \{1,2,...,m\}$ be the set of selected genetic markers. Clearly S is determined by I and λ, so we can write $S = \hat{S}^{\lambda}(I)$. For the entire set of variables $k \subseteq \{1,2,...,m\}$, the probability of variables being selected is defined as

$$\hat{\Pi}_k^{\lambda} = P(k \subseteq \hat{S}^{\lambda}(I))$$

For a cut-off $0 < \pi_{thr} < 1$ and a set of regularization parameters Λ, the set of stable variables is defined as $\hat{S}^{stable} = \{k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}$.

Note by the AIC selection, the parameter set Λ equals to a single point of λ that corresponds to the smallest AIC value. Let N ($|N| < m$) be the set of unrelated variables. Define V to be the number of falsely selected variables with stability selection, then $V = | N \cap \hat{S}^{stable} |$.

The detailed procedure of hybrid AIC and stability selection is described as below.

The hybrid AIC and stability selection algorithm

Step 1: Find the optimal regularization parameter λ that yields the smallest AIC value.

Step 2: Bootstrap the original data *T* times, with subsample size

Step 3: Calculate the selection probability

For *t* from 1 to *T*

Let I_t be the random subsample of $\{1,2,...,n\}$, $|I_t| = \lfloor n/2 \rfloor$

For *k* from 1 to *K*

$$\hat{\Pi}_k^{\lambda} = \frac{\sum_{t=1}^{T} I(\{k\} \subseteq \hat{S}^{\lambda}(I_t))}{T}$$, where I is the indicator function

Step 4: Calculate the cut-off parameter π_{thr} by the pre-set false positive rate

$$\pi_{thr} = \frac{1}{2}(\frac{q_{\Lambda}^2}{m^2} + 1)$$

Step 5: Get the stable variable set $\hat{S}^{stable} = \{k : \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}$

Despite the hybrid AIC and stability selection procedure, we also provide an adaptive method to make the solution to have predetermined *k*-sparsity.

Unified *L_p* algorithm with predetermined sparsity

Step 1: Data normalization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…*m*

Step 2: For any predetermined sparsity level *k*, set iterative index *r*=0, ε =0.0001; Initialize $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$, $\boldsymbol{\beta}^{(0)} = \mathbf{0}$

Step 3: Set $\mathbf{B}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \mu \mathbf{X}^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha}^{(r)} - \mathbf{X}\boldsymbol{\beta}^{(r)})$, and denote $[|\mathbf{B}^{(r)}|]_k$ to be the *k*-th largest element of $|\mathbf{B}^{(r)}|$.

Step 4: Update $\lambda^{(r)} = \frac{\sqrt{96}}{9\mu}([|\mathbf{B}^{(r)}|]_k)^{\frac{3}{2}}$.

Step 5: Update $\boldsymbol{\alpha}^{(r+1)} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(r)})$

Update $\boldsymbol{\beta}^{(r+1)} = R_{\lambda^{(r)}\mu,1/2}(\boldsymbol{\beta}^{(r)} + \mu \mathbf{X}^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha}^{(r+1)} - \mathbf{X}\boldsymbol{\beta}^{(r)}))$

Step 6: Apply the lower bounds to regularize $\boldsymbol{\beta}^{(r+1)}$ and use the SCG algorithm (Zhang *et al.*, 2009) with the initial point $\boldsymbol{\beta}^{(r+1)}$ to find the minimizer $\boldsymbol{\beta}_p^{(r+1)}$ of objective function (2.3)

```
Step 7: Calculate $\| \boldsymbol{\beta}_p^{(r+1)} - \boldsymbol{\beta}_p^{(r)} \|_2$
If $\| \boldsymbol{\beta}_p^{(r+1)} - \boldsymbol{\beta}_p^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step3
```

The $\boldsymbol{\beta}_p^{(r+1)}$ is the final outcome with *k*-sparsity.

2.3 Simulations

In this section, we empirically compared our USR algorithm with single marker test (χ^2 test), Elastic-net, Orthogonal Matching Pursuit (OMP), FOcal Underdetermined System Solver (FOCUSS) (Rao and Kreutz-Delgado 1999) and Random Forest (Goldstein, Hubbard et al. 2010, Chen and Ishwaran 2012). We first compared these algorithms under our own simulation design with and without family structure. In addition, we compared the algorithms under the simulated data from GAW17.

2.3.1 Simulation I: unrelated individuals

To validate our USR, we performed simulation experiments based on the Encyclopedia of DNA Elements (ENCODE) data. This data set contains 522 haplotypes and 1688 SNPs. We used this haplotype pool to generate genotypes and the corresponding phenotypes, i.e., X and Y respectively in the linear mixed-effect model (1). The phenotypes are simulated based on the linear model with assigned causal SNPs under the controlled heritability. We implemented $L_{0.1}$, $L_{0.9}$, $L_{0.5}$, Elastic-net, OMP and FOCUSS methods respectively. The Elastic-net and weighted Elastic-net programs were developed according to Friedman's papers (Friedman, Hastie et al. 2007, Friedman, Hastie et al. 2010); OMP and FOCUSS (Rao and Kreutz-Delgado 1999) programs were downloaded from the link in their publications.

In this simulation, we generated 1000 samples and give the weight for each marker as follows, $weight = 2\sqrt{MAF(1-MAF)}$ where MAF is the minor allele frequency.

The detailed procedure of our experiment is as follows:

Step1: Set the risk haplotype ratio to be 25% (risk haplotypes/all haplotypes); set the iterative index *k*=0, I(0)=Ø.

Step2: *k*=*k*+1; randomly select a SNP as causal variant C(*k*); count the index of the haplotypes that contain C(*k*), and denote this index set as I(*k*) (risk haplotypes);

Step3: I(*k*)=I(*k*-1) $\cup$ I(*k*); if I(*k*)>0.25, jump to Step4, otherwise return to Step2.

Step4: Generate 10000 genotype samples from the pool randomly.

Step5: Calculate each sample’s genetic score S, i.e., how many risk haplotypes that this sample has; S=0,1,2.

Step6: Generate each sample’s phenotype: y=b*S+ε, ε ～ N(0,1), b=sqrt(0.01/(0.99*var(S))).

To evaluate our methods, we compared them with the single marker test (χ^2 test), Elastic-net, OMP and FOCUSS respectively. We also extended our family adjustment and weighted model to the one with Elastic-net penalty. For the numerical algorithm, we used the cyclical coordinate descent, computed along a regularization path.

The receiver operating characteristic (ROC) curve is shown in Figure 2.1.

In this dissertation, the TPR is defined by the number of selected true variants divided by the total number of true variants; and the FPR is defined by the number of selected false variants divided by the total number of false variants.

From Figure 2.1, by calculating the AUR (area under the ROC curve), we can conclude that weighted models with the use of $L_{0.5}$ and $L_{0.1}$ regularization term perform best among all the methods listed above, and the classic single marker test (χ^2 test) has the lowest power. The FPR and TPR of FOCUSS and OMP methods were stuck in a low range, which is difficult to perform a comparison of AUR. In addition, FOCUSS became unstable with the tuning parameter getting larger and its TPR decrease with FPR increase. For the sake of stability and efficiency, we did not perform OMP and FOCUSS methods in the following sections.

Table 2.1. The error rate of using optimal λ selected by the AIC

N=1000,H^2=0.05	TPR	FPR
Elastic-net	0.0745	0.0151
$L_{0.5}$	0.0805	0.0213
$L_{0.1}$	0.0021	1.5883e-4
$L_{0.9}$	0.0018	1.2202e-4

Table 2.1 and Table 2.2 are generated by the average of 100 replicate simulations with 1000 samples and 0.05 heritability. Apparently, the best method should have the highest TPR, while lowest FPR. However, there always exists a trade-off between TPR and FPR.

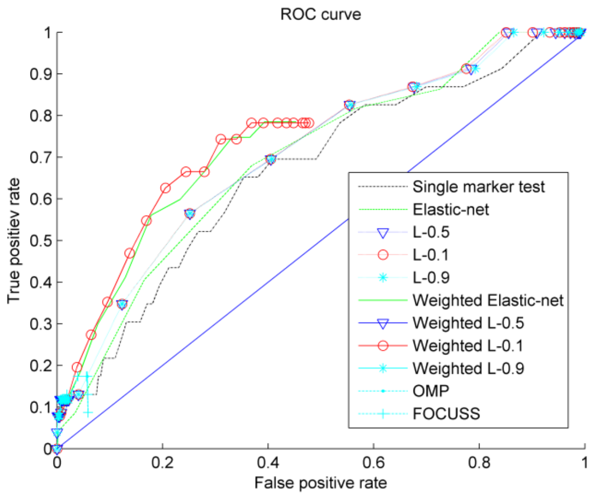

Figure 2.1. Methods comparison under population design of 1000 unrelated individuals. Each point (FPR,TPR) corresponds to a specific λ value, where FPR is false positive rate, and TPR is true positive rate. The large circles stand for the optimal λ selected by the AIC**.**

Table 2.2. The error rate of variables by the hybrid AIC and Stability Selection method

N=1000,H^2=0.05	TPR	FPR
Elastic-net	0.0729	0.0142
$L_{0.5}$	0.0818	0.0208
$L_{0.1}$	0.0337	2.0743e-3
$L_{0.9}$	0.0261	2.3173e-3

By comparing Table 2.1 with Table 2.2, we find that $L_{0,5}$ and Elastic-net had quite similar performance under both AIC and Stability Selection. Under AIC, $L_{0.1}$ and $L_{0.9}$ appeared to be too conservative and yielded extremely low FPR and TPR. However, the stability selection rectified the conservativeness to make corresponding FPR closer to the pre-set type I error threshold (0.05). Therefore, we recommend hybrid stability selection with AIC as a better choice and just present results of using hybrid AIC and Stability Selection in the following sections. Furthermore, the $L_{0.9}$ was shown not as good as $L_{0.1}$ and $L_{0.5}$, which is also supported by the Section **2.3.3**. So we mainly focused on $L_{0.1}$ and $L_{0.5}$ regularization methods for the remaining of the dissertation.

2.3.2 Simulation II: admixed families

We downloaded the genotype data of region ENr113.4q26 from the ENCODE project Consortium. We inferred 180 CEU (Centre d'Etude du Polymorphisme Humain in Utah, USA) and 180 YRI (Yoruban in Ibadan, Nigeria) haplotypes. We observed 1,693 SNPs in total. At each SNP, we chose the minor allele in the YRI haplotype data as the reference allele. Following previous association study on African Americans (Qin, Morris et al. 2010), we adopted $\omega = 0.8$ vs. $\varpi = 0.2$ as YRI-CEU admixture weights. To 'genotype' one admixed subject in the ENr113.4q26 region, we randomly chose one and another haplotype from the YRI or CEU haplotype data sets with probabilities ω vs. ϖ . In this simulated admixture, the frequencies of reference alleles at the 1,693 SNPs range from 0.0011 to 0.5722. This simulation design includes three major steps.

Step1. Generate parental dataset

For each family, we generated father and mother independently. Each subject is composed of two haplotypes; each time we have 80% chance of randomly selecting a haplotype from YRI, and 20% from CEU. The local ancestry $a_i \in \{0,1,2\}$ for the *i*th subject is the number of haplotypes from YRI data. This design does not model recombination in the small region (ENr113.4q26).

Step2. Generate nuclear family with two children

Two children are generated for each family. To generate one child, we randomly selected one haplotype from father and the other from mother. We simulated *N*(=200) families with the same family structure, which is composed of two parents with two children.

Step3. Generate trait values

To be explicit, for each person, we use the following model to generate trait values.

$$Y_i = b\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i\,,\ (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.6}$$

where $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_N)$

In our simulation, the covariate matrix for each family is $\boldsymbol{\Sigma}_j = \frac{2}{3}\boldsymbol{\Phi} + \frac{1}{3}\mathbf{I}$

where $\boldsymbol{\Phi} = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$

is the kinship matrix. To simulate heritability $H^2 = 0.05$, the true model we used is formula (2.6)

where $b = \dfrac{H}{\sqrt{(1-H^2)Var(\mathbf{X}\boldsymbol{\beta})}}$

Using this model, we mainly compared the results with and without pedigree adjustment. Thus, we evenly assigned causal variants to include both rare and common variants and exclude the influence of weighted method. In this simulation, we did not consider any prior knowledge and set all the weight coefficients to be one.

First, we compared our USR algorithm with other feature selection methods (e.g., Random Forest), using the genuine family structure. Second, to illustrate the capability of the USR to adjust for cryptic relatedness, we inferred the kinship matrix using the REAP

(Thornton, Tang et al. 2012) and adopted the inferred kinship matrix when applying our USR.

Figure 2.2(a) shows the comparison among several methods without adjusting for family structure. For the Random Forest, we ranked the variables by their importance factors, and then selected different number of variables. Finally, we drew the corresponding ROC. The ROC and AUR indicate that $L_{0.1}$ is the best method and the single marker test performs the worst.

In Figure 2.2(b), adjusting genuine relatedness and estimated relatedness outperformed the ordinary regression, which ignoring the relatedness. The result of the USR using estimated kinship matrix is close to that of the USR using genuine kinship matrix. Hence, the estimated kinship based USR method is reliable for cryptic relatedness data analysis. The adjustment of a real kinship matrix appeared a bit better. The ROC also indicates that when the FPR, or type I error is low, the $L_{0.1}$ solution is the best choice; when the FPR is higher, the $L_{0.5}$ solution is the best choice. The Elastic-net sparse representation falls in between, indicating that is a more stable solution.

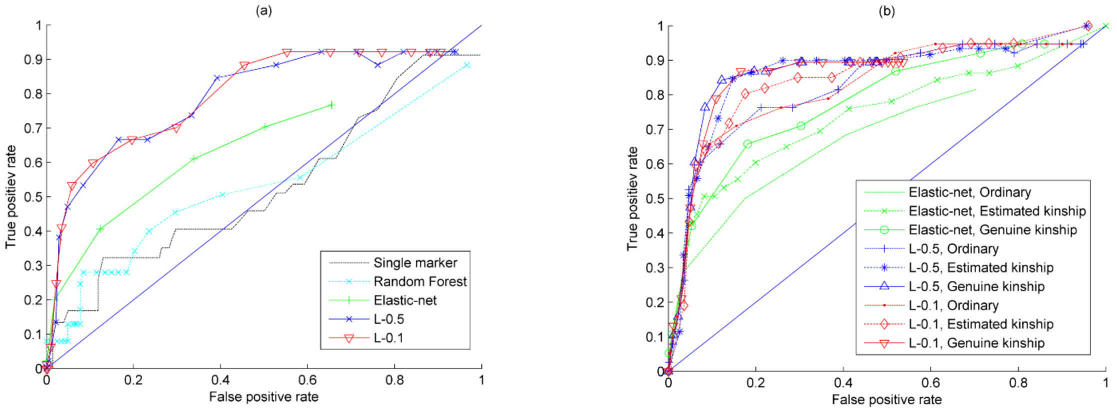

Figure 2.2 Methods comparison under family design of 200 unrelated nuclear families. **(a)** The ROC curves of five methods without adjusting for relatedness.

(b) The ROC curves of two methods with adjusting for relatedness (by estimated kinship matrix and genuine kinship matrix) vs. the ordinary method without adjusting for relatedness.

Table 2.3. The error rate of variables selected by hybrid AIC and Stability Selection

N=800,H^2=0.05	TPR	FPR	AUR
Elastic-net, ordinary	0.0811	0.0223	0.6693
Elastic-net, estimated kinship	0.1291	0.0237	0.6964
Elastic-net, genuine kinship	0.1351	0.0243	0.7271
$L_{0.5}$, ordinary	0.0811	0.0169	0.7789
$L_{0.5}$, estimated kinship	0.2320	0.0201	0.7808
$L_{0.5}$, genuine kinship	0.2432	0.0205	0.8126
$L_{0.1}$, ordinary	0.0435	3.623e-3	0.7884
$L_{0.1}$, estimated kinship	0.0501	4.521e-3	0.8063
$L_{0.1}$, genuine kinship	0.0526	4.753e-3	0.8198

Table 2.3 is generated by the average of 100 replicate simulations with 200 nuclear families (800 samples) and 0.05 heritability. In terms of AUR, the $L_{0.1}$ and $L_{0.5}$ Family

adjustment models are the best models. The result confirms again that the model with adjusted family structure yields higher TPR while lower FPR and FDR (false discover rate).

2.3.3 Comparison of different L_p norm

Here we show the comparison of different L_p norm regularization model based on Simulation II: the admix family design. In this simulation, we change *p* from 0.1 to 0.9 and compare their AUR with Elastic-net regularization and single marker test. Similar to the design in Simulation II, we set the total heritability to be 5% and repeat each simulation 500 times.

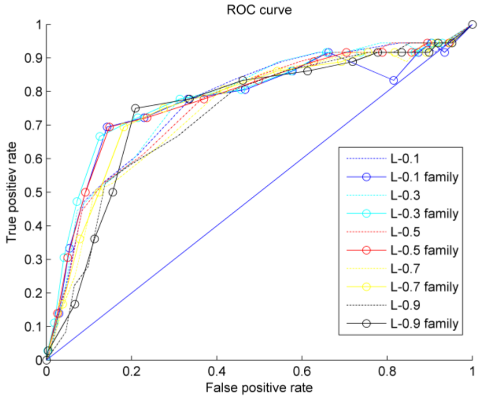

Figure 2.3. The ROC curve comparison for six different L_p norms under admixed family design with 23 rare causal variants. The legend "Family" stands for regression with adjustment of family structure.

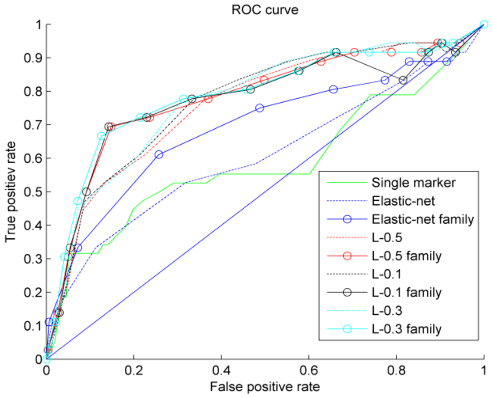

Figure 2.4. The ROC curve comparison for three different L_p norms, Elastic-net and single marker test under admixed family design with 23 rare causal variants.

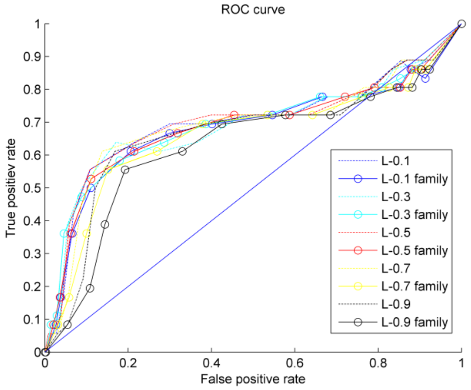

Figure 2.5. The comparison of ROC curve for six different L_p norms under admixed family design with 23 rare and 15 common causal variants. The legend "Family" stands for regression with adjustment of family structure.

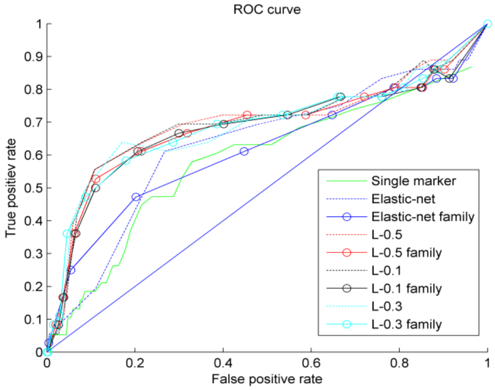

Figure 2.6. The comparison of ROC curve for three different L_p norms, Elastic-net and single marker test under admixed family design with 23 rare and 15 common causal variants.

Table 2.4. The AUR comparison under different methods with 23 rare causal variants, where AUR_F represents regression with adjustment of family structure

Methods	AUR	AUR_F
L_p (p=0.1)	0.788354	0.812526
L_p (p=0.2)	0.789371	0.814508

Table 2.5. The AUR comparison under different methods with 23 rare and 15 common causal variants, where AUR_F represents regression with adjustment of family structure

Methods	AUR	AUR_F
L_p (p=0.1)	0.704866	0.711651
L_p (p=0.2)	0.70587	0.71215
L_p (p=0.3)	0.706421	0.713496

L_p (p=0.3)	0.79092	0.821876
L_p (p=0.4)	0.786989	0.820463
L_p (p=0.5)	0.778918	0.819813
L_p (p=0.6)	0.777636	0.817066
L_p (p=0.7)	0.769813	0.813395
L_p (p=0.8)	0.76676	0.809379
L_p (p=0.9)	0.762718	0.799369
Elastic-net	0.669337	0.72705
Random Forest	0.590643	
single marker test	0.614267	

L_p (p=0.4)	0.706007	0.714173
L_p (p=0.5)	0.704962	0.712474
L_p (p=0.6)	0.700806	0.707429
L_p (p=0.7)	0.694789	0.700064
L_p (p=0.8)	0.690193	0.692156
L_p (p=0.9)	0.681642	0.683268
Elastic-net	0.643623	0.6598
Random Forest	0.575592	
single marker test	0.549118	

Figure. 2.3 to 2.6 and Table 2.4, Table 2.5 indicate that when p<0.5, the L_p norm based regularization yeilds the highest AUR, while any L_p norm based regularization outperforms the Elastic-net based test. The single marker test always gives the worse result.

2.4 GAW18 data analysis

To further demonstrate the effectiveness of our USR, we compared it with competitors under an official simulation from the Genetic Analysis Workshop 17 (GAW17). This data set contains real genotypes of 24,487 SNPs from 3,205 genes on 697 subjects, together with simulated phenotypes of these subjects. We chose replicate 1 of Q1 as outcomes and applied the algorithms to locate promising SNPs from genotype data. Both the weighted and unweighted versions of our USR detected 5 causal SNPs within two genes (FLT1 and KDR). Three of the causal SNPs were rare variants but were missed by the single marker test (Figure 2.7, Table 2.6). Again, in this comparison, our USR inclined to discover rare casual variants with higher true positive than the single marker test.

To illustrate effectiveness of our algorithm to locate rare genetic variants, we applied it to the analysis of Mexican Americans sequence data from the GAW18. This dataset contains next generation sequencing data of 850 subjects within 21 large families.

2.4.1 Simulated phenotype analysis

First, we analyzed the simulated DBP (diastolic blood pressure), where DBP was set to be influenced by 1243 variants of 245 genes and 1040 variants of 205 genes respectively. After quality control, we selected 504 subjects within the region of 1244 variants from three genes (SLC35E2, TNN and MAP4) that influenced the phenotypic data. All the data we used were from the first visit of the longitudinal data. We connected raw DBP data with covariates and population structure (adjusted by the first 10 PCs of the genotypic matrix) and pedigree structure by our generalized sparse regression model.

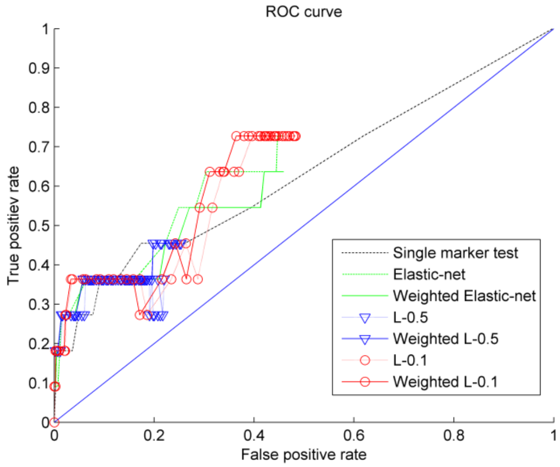

Figure 2.7. The ROC of Chromosome 13 data. The weights are generated by the correlation coefficients between phenotypes and variants.

Table 2.6. Identified casual rare variants for phenotype

Causal gene and SNPs	Single marker test	Elastic-Net	Weighted Elastic-Net	$L_{0.5}$	Weighted $L_{0.5}$	$L_{0.1}$	Weighted $L_{0.1}$	MAF
KDR/C4S1874	×	√	√	√	√	√	√	0.00717

KDR/C4S1877	√	√	√	√	√	√	√	0.164993
KDR/C4S1884	×	×	√	√	√	√	√	0.0208
KDR/C4S1887	×	×	×	√	√	√	√	0.00717
FLT1/C13S523	√	√	√	√	√	√	√	0.066714
FLT1/C13S523	×	×	√	√	√	√	√	0.004304

Note: A "√" indicated that the corresponding marker was detected as a causal marker by a particular method. A × indicated that that the corresponding marker was not detected as causal marker by a particular method.

Table 2.7. The error rate of GAW18 data

N=504, SNPs=1243	TPR	FPR	AUR
Elastic-net	0.173913	0.070434	0.643622
Elastic-net Family	0.26087	0.160524	0.678642
Elastic-net Family&Weight	0.217391	0.05733	0.681642
$L_{0.5}$	0.173913	0.045864	0.694789
$L_{0.5}$ Family	0.217391	0.052416	0.707428
$L_{0.5}$ Family&Weight	0.217391	0.074529	0.712473
$L_{0.1}$	0.26087	0.09828	0.700806
$L_{0.1}$ Family	0.217391	0.052416	0.711650
$L_{0.1}$ Family&Weight	0.26087	0.09828	0.714172

In this analysis, our USR algorithm appeared to have better TPR and better AUR compared to the algorithms without adjusting pedigree structure, while maintaining almost the same FPR level (Table 2.7). The pedigree adjustment appeared to be both necessary and beneficial as shown by this set of results.

2.4.2 Real phenotype analysis

Finally, we applied the proposed USR to analyze real DNA sequence data on DBP and SBP (systolic blood pressure) from GAW18. After quality control, we obtained GWAS data of 783 Mexican Americans with 438,790 SNPs and NGS data of 506 Mexicans with 6,824,165 SNPs. When analyzing the GWAS data using our USR algorithm, we obtained sparse representation for each chromosome by choosing the entire chromosome as a window. However, for the sequence data set, it is too large to be analyzed as a whole window. Thus, we divided each of the large chromosome (1,3,5,7,9) into two equal parts and obtained their sparse representations separately.

Based on above algorithm, we analyzed GWAS and sequence data by our USR separately to find the susceptible genetic variants. Combining the significant variants selected by both GWAS and sequence data, we identified 23 promising genes (Table 3S). We also identified 3 significant pathways relevant to hypertension by pathway wise SKAT (Wu, Lee et al. 2011). The most significant pathway (p=3.24e-8) was Glioma, including BRAF, SHC3, CAMK2B, EGFR, and PDGFRB. An independent study (Houben, Louwman et al. 2004) suggested that Glima pathway would be associated with hypertension through potentially neurocarcinogenic effects of antihypertensive medication. The second most significant pathway (p=2.74e-7) is the regulation of actin cytoskeleton pathway, including GNA12, BRAF, EGFR, PDGFRB and PIP5K1B. This pathway was identified to be associated with hypertension by an independent study (Tripodi, Valtorta et al. 1996). The third most significant pathway (p=3.87e-6) is chronic myeloid leukemia pathway, including BRAF, RUNX1, SHC3 and MECOM. This pathway, as suggested by independent studies (Guymer, Cairns et al. 1993, Dumitrescu, Seck et al. 2011), would highly influence benign intracranial hypertension and pulmonary arterial hypertension.

Furthermore, we found some new candidate genes and pathways that were not reported in the previous independent study. For example, FMO1 (p=9.81e-5) is a risk gene of cardiovascular disease (Mendelsohn and Larrick 2013), which is usually associated with hypertension. Another susceptible gene is AGBL1 (p=8.16e-4), which is associated with carotid plaque (Dong, Beecham et al. 2012), and prehypertension is associated with significantly increased carotid atherosclerotic plaque (Hong, Wang et al. 2013). We also report long-term depression pathway (p=4.21e-6) as a significant pathway. It might cause depression that is a risk factor of hypertension (Meng, Chen et al. 2012).

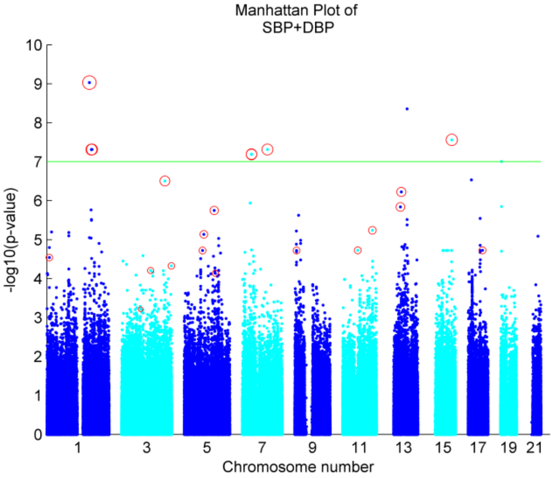

Figure 2.8. The Manhattan plot for SNPs on odd numbered chromosomes. The p-values were computed from single marker tests. The red circles stand for the markers selected by our USR. We used SBP+DBP as the phenotype. The genome-wide nominal significance level was set to be 10^{-7}, as shown by the green horizontal line.

2.5 Conclusion

Many existent sparse regression algorithms assume unrelated subjects. Such algorithms fail to adjust for complex pedigrees and cryptic relatedness as often occur in the genomic data. In this article, we have proposed the USR algorithm for variant selection from DNA sequence data with an arbitrary intra-individual relationship and population structure. Our USR algorithm allows informative weighting to incorporate prior knowledge. This approach provides a flexible way to adjust for preference or risk variants. Extensive simulation results indicated that a properly predetermined weighting scheme can notably improve selection accuracy of causal variants.

Our algorithm can handle both rare and common variants with equal efficiency. The ability of our algorithm to pinpoint causal variants, especially rare causal variants, was clearly demonstrated by intensive simulations. We suggest using L_p norms ($0.1< p <0.5$) in the model since these regularization terms provide better performance in terms of AUR, TPR and FPR. For the sake of computational speed, $L_{0.5}$ norm is a better choice. In particular, our algorithm can solve the low sample size but high dimensional feature problem, that is, sample size is less than the number of variants, as often happens in genomic studies.

Like existent methods, our algorithm has some limitations. First, it focuses on a single variant effect on a trait of interest. A more powerful strategy would be to group multiple variants and incorporate group wide information into the model. Doing so, however, would scarify single marker resolution. Second, our algorithm assumes linear relationship between phenotype and genotype, which may be unrealistic for many scenarios in practice. Extension to nonlinear regression models call for additional efforts. Last, it deserves further investigation on how to choose the optimal tuning parameter and the optimal set of features.

CHAPTER 3 SIGNIFICANCE TESTS FOR UNIFIED SPARSE REGRESSION

3.1 Introduction

Deep sequencing technologies have been generating huge amounts of data of rare and common DNA sequence variants. A number of sequence association tests have been developed for identifying marker sets (e.g., a group of SNPs or CNVs) that contain functional genetic variants. Most of them, however, do not jointly model cryptic relatedness, population structure and other covariates. With the growing demand of analyzing next generation sequencing data of multi-ethnic individuals, linear mixed models have become popular because of their demonstrated effectiveness in accounting for sample relatedness (Amos 1994) and population structure which occurs when there are large-scale systematic differences in genetic ancestry among individuals in a sample. Typical examples include individuals with various levels of immigrant ancestry and more recent shared ancestors than one would expect in a homogenies population. Cryptic relatedness, refers to the presence of relatives in a sample of ostensibly unrelated individuals, could pose more serious confounding than population structure (Devlin and Roeder 1999), especially for samples from small and isolated populations (Voight and Pritchard 2005). Accounting for population structure is more challenging when family structure or cryptic relatedness is also present (Price, Zaitlen et al. 2010). We paved the way to correct for the effects of both confounders jointly.

Within the framework of linear mixed models, famSKAT (Chen, Meigs et al. 2013) and GEMMA (Zhou and Stephens 2012) appeared as two powerful sequence association tests for identifying small marker sets that harbor dense functional genetic variants. FamSKAT is a set based test which is an extension of SKAT to be applicable to family data. GEMMA is a computationally efficient method for fitting multivariate linear mixed models. These prominent tests require that the number of markers in a testing set is much smaller than the sample size. However, in population deep sequencing studies, one encounters quite often high dimensional data sets (HDS), where the number of marker loci is larger than the sample size and the number of functional variants is very small. The aforementioned tests are incapable of identifying the functional variants on such sparse HDS. With high-dimensional sparse functional marker data sets, the aforesaid tests are incapable to identify them. Some sparse regression methods were developed to localize individual functional markers from high-dimensional marker sets, jointly modeling pedigree structure and population structure. They include Lasso (Rakitsch, Lippert et al. 2013), Ridge regression (Endelman 2011), Elastic-net (Zou and Hastie 2005) and the USR that we proposed recently (Cao, Qin et al. 2014). However, these methods yield biased solutions and are ineffective to prevent false discoveries of random markers and high-dimensional marker sets irrelevant to functional variants.

In this article, we first present a unified test (uFineMap) for accurately localizing causal loci. The uFineMap is a marker wise test under a scaled sparse linear mixed regression, which jointly models marker wise effect, relatedness and population stratification. It applies scaled L_p ($0 < p <1$) norm regularization to generate a de-biased solution. Next, we present an additional significant test (uHDSet) for identifying high-dimensional sparse associations in deep sequencing genomic data of related individuals. The uHDset integrates the marker wise statistics of the uFineMap to identify susceptible high-dimensional marker sets. In the uHDSet, the dependence among markers is modeled to appropriately control set-based Type I error rates. Under extensive simulations, the uFineMap outperformed the

GEMMA (Zhou and Stephens 2012) and a Scaled Lasso based method (Javanmard and Montanari 2014). The uHDSet yields higher statistical power than famSKAT and GEMMA. Applications to Framingham Heart Study also show that our method yields novel interesting candidate genes and pathways for follow-up studies, showing its advantages over the two compared prominent alternative methods. Finally, caveats of the proposed methods and perspective future efforts are discussed.

3.2 Unified scaled L_p norm regularized regression model

We still assume that the phenotypes, genotypes and covariates follow eq. (2.2). To avoid the puzzle of tuning parameter selection and reduce the uncertainty of sparse regression methods for model selection, we adopt the idea of scaled sparse linear regression(Sun and Zhang), which jointly estimates the regression coefficients and the noise level of the data. The estimated noise level is critical to correct the bias caused by the regularization term. With a correction of the bias, the de-biased estimator is applied to construct uFineMap statistics for each variable before testing for marker wise significance for each variant.

For the scaled L_p norm based sparse regression problem, we modify the problem to the following form:

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}) = \underset{\boldsymbol{\alpha}\in R^L, \boldsymbol{\beta}\in R^m, \sigma>0}{\arg\min} \{ \frac{(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})}{2n\sigma} + \frac{\sigma}{2} + \lambda \| \boldsymbol{\beta} \|_p^p \} \tag{3.1}$$

In the unified scaled sparse regression, the tuning parameter λ is updated in an iterative procedure. But we still need to choose an initial tuning parameter λ_0 to reach a solution. However, the selection of the λ_0 is more flexible and less sensitive to our significance test. Because the estimated noise level $\hat{\sigma}$ and the bias caused by the L_p norm regularization are both proportional to the initial λ_0 , they can be compensated by the procedure of constructing de-biased estimators. Furthermore, scaled *L_p* norm regularization can produce a robust and consistent estimation of the regression coefficients, which is critical for developing the asymptotic distribution of the de-biased estimators. For these reasons, we use scaled sparse regression.

To solve the optimization problem (3.1), we combine the algorithm for unified L_p norm based sparse regression(Cao, Qin et al.) with that for the general scaled sparse regression(Sun and Zhang 2012) and propose the following algorithm.

Algorithm for unified scaled L_p norm sparse regression

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…*m*

Step 2: Initialize $\lambda^{(0)} = \lambda_0 = 2\sqrt{\frac{\log(m)}{n}}$, $\sigma^{(0)} = \sqrt{\frac{\mathbf{Y}^T \mathbf{\Sigma}^{-1}\mathbf{Y}}{n}}$, $\hat{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ set iterative index *r*=0, $\varepsilon = 0.0001$

Step 3: Update $\hat{\sigma}, \lambda, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$ coordinately

$$\hat{\sigma}^{(r+1)} = \sqrt{\frac{1}{n}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\alpha}}^{(r)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(r)})^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\alpha}}^{(r)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(r)})}$$

$$\lambda^{(r+1)} = \hat{\sigma}^{(r+1)}\lambda_0$$

Update regression coefficients by USR algorithm:

$$(\hat{\boldsymbol{\alpha}}^{(r+1)}, \hat{\boldsymbol{\beta}}^{(r+1)}) = \underset{\boldsymbol{\alpha}\in R^L, \boldsymbol{\beta}\in R^m}{\arg\min}\{\frac{(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})}{2n\hat{\sigma}^{(r+1)}} + \lambda^{(r+1)} \| \boldsymbol{\beta} \|_p^p\}$$

Step4: If $\| \boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step 3

3.3 The de-biased version of scaled L_p norm regularized sparse regression

In HDSs (high dimensional sets), to obtain a stable and sparse solution, a regularization term is often needed. Take two widely used methods, Lasso and Ridge regression, as examples. Both of them utilize the regularization term to assure a unique and stable solution. On one hand, the regularization term can enforce most regression coefficients to shrink exactly to zero, contributing to dimension reduction; on the other hand, the bias introduced by the regularization makes the estimated non-zero regression coefficients inclined to be smaller than their true values.

To assess the asymptotic Gaussian distribution of sparse regression coefficients, a de-biased estimator is constructed. Adopting the idea of unbiased estimation(Bühlmann 2013, Javanmard and Montanari), we develop a de-biased estimator to recover the original unbiased regression coefficients. The detailed algorithm procedure is presented below.

The Algorithm for de-biased estimator

Step 1: Set $\gamma = \frac{\hat{\lambda}}{\hat{\sigma}}$, where $\hat{\lambda}$ and $\hat{\sigma}$ is the estimated parameters of the scaled sparse regression (3.1)

Step 2: Set $\mathbf{Z} = (\mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{X}) / n$

Step 3: For *i*=1,2,…,*m*, solve z_i by the following constraint convex program:

$$\begin{aligned} &\text{minimize} && \mathbf{m}^T \mathbf{Z} \mathbf{m} \\ &\text{subject to} && \| \mathbf{Zm} - \mathbf{e}_{\mathrm{i}} \|_\infty \le \gamma \end{aligned}$$

Step 4: Set $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_m)^T$

If any of the above problems is not feasible, then set $\mathbf{M} = \mathbf{I}_{m \times m}$

Step 5: Define the unbiased estimator by $\hat{\boldsymbol{\beta}}^u = \hat{\boldsymbol{\beta}} + \frac{1}{n} \mathbf{M} \mathbf{X}^T \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}})$ (3.2)

Where $\hat{\boldsymbol{\beta}}$ is the solution of formula (3.1)

3.4 Hypothesis tests and confidence intervals

Without loss of generality, we assume that the following true model holds

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma} = \sigma_\Phi^2\boldsymbol{\Phi} + \sigma_\varepsilon^2\mathbf{I})$$

where $\mathbf{Y}$ is the covariates adjusted phenotype; $\boldsymbol{\beta}_0$ is the ground truth regression coefficients which stands for true signal.

We define the sparsity level of $\boldsymbol{\beta}_0$ as $s_0 = \{i \in \{1,2,...,m\} \mid \beta_{0,i} \neq 0\}$. In this dissertation, we apply a weak assumption for the sparse model, which is $s_0 = O(n/\log(m))$. Without any further notice, we always assume that this assumption holds.

3.4.1 uFineMap test

For each marker *i*, our goal is to develop a significance test to determine whether each regression coefficient β_i is significant or not. For a specific $i \in \{1,2,...,m\}$, we define the null hypothesis H$_0$: $\beta_i = 0$ and the alternative hypothesis H$_1$: $\beta_i \neq 0$

Assuming the linear mixed model (2.2) with Gaussian noise and fixed design matrix ***X*** and considering the de-biased estimator (3.2), we have the following asymptotic distribution.

$$n(\hat{\boldsymbol{\beta}}^{\mu} - \hat{\boldsymbol{\beta}}^{0}) \xrightarrow{d} N(0, \sigma^2\mathbf{M}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T)$$

With this property, we can directly derive the significance test for each marker. The *p*-value for each variable can be defined by the following:

$$P(i) = 2(1-\Phi(\frac{n\,|\,\hat{\beta}_i^{\mu}\,|}{\hat{\sigma}\sqrt{[\mathbf{MX}^T\mathbf{\Sigma}^{-1}\mathbf{XM}^T]_{i,i}}})),\ i = 1,2,...,m$$

where Φ is the cumulative distribution function of a standard normal distribution.

Intuitively, an asymptotic two-sided confidence interval for variable *i* with a significance level α is expressed as below:

$$[\beta_i^{\mu} - \delta(\alpha), \beta_i^{\mu} + \delta(\alpha)],\ \text{where}\ \delta(\alpha) = \Phi^{-1}(1-\alpha/2)\hat{\sigma}n^{-1}\sqrt{[\mathbf{MX}^T\mathbf{\Sigma}^{-1}\mathbf{XM}^T]_{i,i}}$$

Proof:

We first need to define the **Compatibility condition**. For a fixed design matrix, we propose the following compatibility condition.

There exists a $\phi_0 > 0$, such that for any $\mathbf{\beta}$ satisfying $\|\mathbf{\beta}_{S_0^c}\|_p^p \le 3\|\mathbf{\beta}_{S_0}\|_p^p$, the following inequality holds

$$\|\mathbf{\beta}_{S_0}\|_p^{2p} \le \frac{s_0\mathbf{\beta}^T\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X\beta}}{n\phi_0^2} \qquad \text{(S3.1)}$$

In general, it is a relatively mild condition, but it is required to ensure the asymptotic property for the estimators.

The estimator in (3.1) should satisfy Karush-Kuhn-Tucker (KKT) conditions, which implies:

$$-\frac{\mathbf{X}^T\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{\beta}})}{n} + \lambda\hat{\mathbf{d}} = 0,\ \text{where}\ \hat{d}_j = sign(\hat{\beta}_j)\,p\hat{\beta}_j^{1-p}$$

The KKT conditions can be rewritten as

$$\mathbf{Z}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)+\lambda\hat{\mathbf{d}}=\frac{\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}}{n}$$

where $\mathbf{Z}$=($\mathbf{X}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}$)/n

Define $\boldsymbol{\Delta}:=\sqrt{n}(\mathbf{MZ}-\mathbf{I})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)$. It follows that

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0+\mathbf{M}\lambda\hat{\mathbf{d}}=\frac{\mathbf{MX}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}}{n}-\frac{\boldsymbol{\Delta}}{\sqrt{n}}$$

Based on (3.2), it holds that

$$\hat{\boldsymbol{\beta}}^u=\hat{\boldsymbol{\beta}}+\frac{1}{n}\mathbf{MX}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{Y}\text{ - }\mathbf{X}\hat{\boldsymbol{\beta}})\,.$$

The last two equations combined with the KKT conditions imply that

$$(\hat{\boldsymbol{\beta}}^u-\boldsymbol{\beta}^0)=\frac{\mathbf{MX}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}}{n}+\frac{\boldsymbol{\Delta}}{\sqrt{n}} \qquad \text{(S3.2)}$$

Next we will show that $\frac{\boldsymbol{\Delta}}{\sqrt{n}}$ is asymptotically negligible.

$$\begin{aligned}&\|\frac{\boldsymbol{\Delta}}{\sqrt{n}}\|_\infty=\|(\mathbf{MZ}-\mathbf{I})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_\infty\\&\le\|(\mathbf{MZ}-\mathbf{I})\|_\infty\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_\infty \qquad \text{(S3.3)}\\&\le\gamma\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_p\end{aligned}$$

We need to estimate the bound of $\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_p$

According to the basic inequality ((Bühlmann and Van De Geer 2011) Lemma 6.1), it holds that

$$\frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{2n\sigma}+\lambda\|\hat{\boldsymbol{\beta}}\|_p^p\le\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^0)}{2n\sigma}+\lambda\|\boldsymbol{\beta}^0\|_p^p$$

and

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)^T\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)}{2n\sigma}+\lambda\|\hat{\boldsymbol{\beta}}\|_p^p\le\frac{\boldsymbol{\varepsilon}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)}{n\sigma}+\lambda\|\boldsymbol{\beta}^0\|_p^p \qquad \text{(S3.4)}$$

We define $J := \{\max_{1 \le j \le m} | \boldsymbol{\varepsilon}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}^{(j)} | / n\sigma \le \lambda_0\}$

On J , by assuming $\lambda \ge 2\lambda_0$ and using (S3.4), we have

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + 2\lambda \| \hat{\boldsymbol{\beta}} \|_p^p \le \lambda \| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \| + 2\lambda \| \boldsymbol{\beta}^0 \|_p^p$$

Applying the triangle inequality to the left hand side yields

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + 2\lambda (\| \hat{\boldsymbol{\beta}}_{S_0} \|_p^p + \| \hat{\boldsymbol{\beta}}_{S_0^c} \|_p^p) \ge \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + 2\lambda (\| \hat{\boldsymbol{\beta}}_{S_0^c} \|_p^p - \| \hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0 \|_p^p + \| \boldsymbol{\beta}_{S_0}^0 \|_p^p)$$

Similarly, on the right-hand side we have

$$\lambda \| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \|_p^p + 2\lambda \| \boldsymbol{\beta}^0 \|_p^p \le \lambda (\| \hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0 \|_p^p + \| \hat{\boldsymbol{\beta}}_{S_0^c} \|_p^p) + 2\lambda \| \boldsymbol{\beta}^0 \|_p^p$$

Now, combining the left- and the right-hand side we conclude that

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + \lambda \| \hat{\boldsymbol{\beta}}_{S_0^c} \|_p^p \le 3\lambda \| \hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0 \|_p^p \tag{S3.5}$$

Inserting the compatibility condition (S3.1) into (S3.5) we get following inequality on J

$$\begin{aligned}
&\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + \lambda \| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \|_p^p \\
&\le \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + \lambda \| \hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0 \|_p^p + \lambda \| \hat{\boldsymbol{\beta}}_{S_0^c} \|_p^p \\
&\le 4\lambda \| \hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0 \|_p^p \\
&\le \frac{4\lambda \sqrt{s_0} \| \boldsymbol{\Sigma}^{-\frac{1}{2}} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \|_2}{\sqrt{n} \phi_0} \\
&\le \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n\sigma} + \frac{4\lambda^2 s_0}{\phi_0^2}
\end{aligned}$$

We insert (S3.5) in the first inequality, and use compatibility condition (S3.1) in second inequality. For the third inequality we use $4uv \le u^2 + 4v^2$

Hence, on J it holds that

$$\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \|_p^p \le \frac{4\lambda s_0}{\phi_0^2} \qquad \text{(S3.6)}$$

Combining the result of (S3.3) and (S3.6), we come up with: $\| \frac{\boldsymbol{\Delta}}{\sqrt{n}} \|_\infty \le \gamma(\frac{8\lambda_0 s_0}{\phi_0^2})^{\frac{1}{p}}$ (S3.7)

Moreover, we need to show that the set J and bound (S3.7) is ubiquitous.

Suppose that all the diagonal elements of the covariance matrix $\frac{\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}}{n}$ equal to one (we can always scale the covariance matrix). Then we have $V_j := | \boldsymbol{\varepsilon}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}^{(j)} | / \sqrt{n\sigma^2}$ is $N(0,1)$-distributed.

For any $t \in \mathbb{R}$, define $\lambda_0 := \sqrt{\frac{t^2 + 2\log m}{n}}$, we have following inequality hold:

$$\begin{aligned} & P(J = \{\max_{1 \le j \le m} | \boldsymbol{\varepsilon}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}^{(j)} | / n\sigma \le \lambda_0 \}) = 1 - P(\max_{1 \le j \le m} | V_j | > \sqrt{t^2 + 2\log m}) \\ & \ge 1 - 2m \exp(-\frac{t^2 + 2\log m}{2}) = 1 - 2\exp(-\frac{t^2}{2}) \end{aligned} \qquad \text{(S3.8)}$$

From (S3.7) and (S3.8), we come up with the following conclusion:

$$\begin{aligned} & P(\| \frac{\boldsymbol{\Delta}}{\sqrt{n}} \|_\infty > \gamma(\frac{8 s_0 \sqrt{t^2 + 2\log(m)}}{\sqrt{n}\phi_0^2})^{\frac{1}{p}}) = 1 - P(\| \frac{\boldsymbol{\Delta}}{\sqrt{n}} \|_\infty \le \gamma(\frac{8 s_0 \sqrt{t^2 + 2\log(m)}}{\sqrt{n}\phi_0^2})^{\frac{1}{p}}) \\ & \le 2\exp(-\frac{t^2}{2}) \end{aligned}$$

So, $\| \boldsymbol{\Delta} \|_\infty = Op(s_0 \sqrt{\log(m)} / \sqrt{n}) = op(1)$ (S3.9)

Finally, from (S3.2) and (S3.9), we conclude that

$$n(\hat{\boldsymbol{\beta}}^u \text{-} \boldsymbol{\beta}^0) \xrightarrow{d} N(0, \sigma^2 \mathbf{M}\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\mathbf{M}^T) \qquad \square$$

3.4.2 uHDSet test

Starting from the uFineMap test, we still need to control the family-wise error rate (FWER), i.e., the Type I error rates to claim the whole significant region. Similar to the classical region based test, there are some commonly used correction methods based on p-values from uFineMap tests to control the FWER or false discovery rate (Benjamini and Hochberg 1995, Benjamini and Hochberg 2000). We intend to develop an efficient multiple testing adjustment, taking dependence into consideration, which would be more powerful than uncorrelated adjustment (e.g. the Bonferroni-Holm correction).

For uHDSet test, the null hypothesis is H0: $\beta_1 = \beta_2 = ... = \beta_m = 0$, and the alternative hypothesis H1: $\exists \beta_i \neq 0, i \in \{1,2,..., m\}$

We borrowed the idea of van de Geer et al. (van de Geer, Bühlmann et al. 2013) to construct a new statistic for uHDSet significance test by utilizing the previous uFineMap tests: $S = \max_{i\in\{1,2,...,m\}} \frac{n\,|\,\hat{\beta}_i^{\mu}\,|}{\hat{\sigma}\sqrt{[\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T]_{i,i}}}$. For an arbitrary $z \in R$, the following equation holds.

$$P(S \le z \,|\, \mathbf{X}) - P(\max_{i\in\{1,2,...,m\}} \frac{|\,U_i\,|}{\hat{\sigma}\sqrt{[\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T]_{i,i}}} \le z \,|\, \mathbf{X})) \to 0$$

where $\mathbf{U} \sim N(\mathbf{0}, \sigma^2\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T)$

Under null hypothesis H0, $\max_{i\in\{1,2,...,m\}} \frac{U_i^2}{\hat{\sigma}^2\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T}$ is asymptotically equivalent to the maximum of a sequence of dependent $\chi^2(1)$ variables, whose distribution relies on the design matrix $\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}$. For a fixed matrix $\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}$, we can easily simulate its distribution and use its quantile to estimate the p-value of the uHDSet statistic *S*.

Proof:

From the marker wise theory, for any $i \in \{1, 2, ..., m\}$, we have

$$\hat{\beta}_i^u - \beta_i^0 \xrightarrow{d} N(0, \sigma^2[\mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T]_{i,i} / n)$$

$$P(\frac{\sqrt{n} \,|\, \hat{\beta}_i^u - \beta_i^0 \,|}{\hat{\sigma}\sqrt{[\mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T]_{i,i}}} \le z \,|\, \mathbf{X}) - \Phi(z) \to 0$$

Define: $\mathbf{W} \sim N(0, \hat{\sigma}^2 \mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T)$

Thus $P(\frac{\sqrt{n} \,|\, \hat{\beta}_i^u - \beta_i^0 \,|}{\hat{\sigma}\sqrt{[\mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T]_{i,i}}} \le z \,|\, \mathbf{X}) - P(\frac{|W_j|}{\hat{\sigma}\sqrt{[\mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T]_{j,j}}} \le z \,|\, \mathbf{X}) \to 0$ holds for any $i, j \in \{1, 2, ..., m\}$

As a result, we come up with

$$P(\max_{i \in \{1,2,...,m\}} \frac{\sqrt{n} \,|\, \hat{\beta}_i^u - \beta_i^0 \,|}{\hat{\sigma}\sqrt{[\mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T]_{i,i}}} \le z \,|\, \mathbf{X}) - P(\max_{i \in \{1,2,...,m\}} \frac{|W_i|}{\hat{\sigma}\sqrt{[\mathbf{MX}^T \mathbf{\Sigma}^{-1} \mathbf{XM}^T]_{i,i}}} \le z \,|\, \mathbf{X}) \to 0$$

3.5 Simulation: Complex Families

To further compare different methods fairly, instead of using our own or over-simplified simulation data, we used the software SeqSIMLA2 (Chung, Tsai et al. 2015). SeqSIMLA2 can simulate sequence data in families under quantitative disease models. Using SeqSIMLA2, we generate quantitative traits for 8 large families with 67 individuals and 1000 SNPs in total.

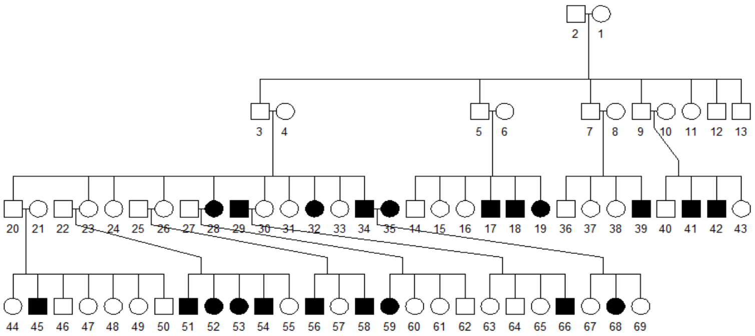

Figure 3.1. The family tree for each simulated family in complex family simulation

3.5.1 Type I error rate evaluation

To verify the validity of our proposed tests, we need to evaluate if the Type I error is well controlled under the null hypothesis. Figure 3.2(a) and 3.2(b) show the Q-Q plots for uFineMap test and uHDSet test respectively. The results indicate that the Type I error rate is appropriately controlled in complex family structure.

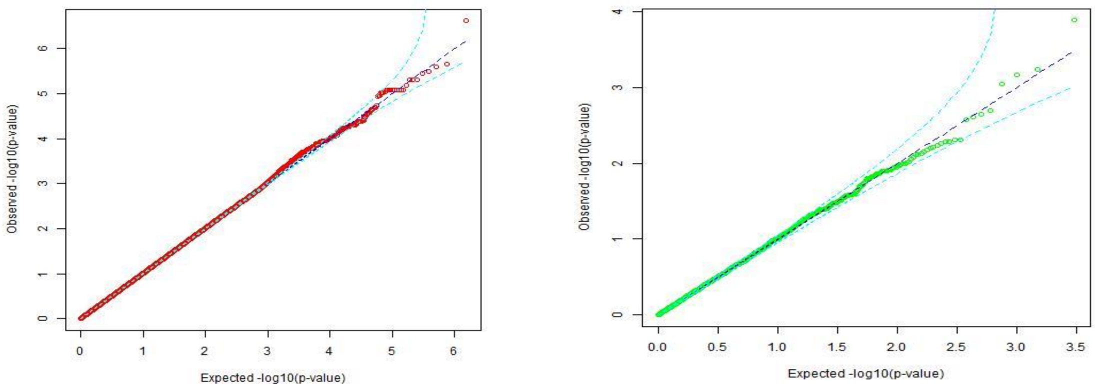

Figure 3.2. **(a)** The Q-Q plot for uFineMap test

(b) The Q-Q plot for uHDSet test

3.5.2 Power comparison

We randomly assign 50 causal variants (25 common, 25 rare) to generate the continuous phenotype. Additionally, we proposed two simulation setting for markers effects. We assign all causal markers to be positively related to the trait value for the same causal direction setting. For the different causal direction setting, half of the causal markers are randomly given a negative relationship with the trait value.

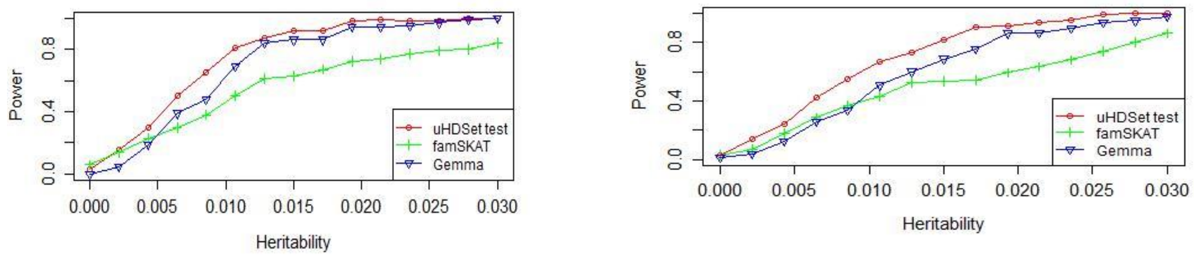

Figure 3.3. (a) Power comparison for uHDSet test with same causal direction.

(b) Power comparison for uHDSet test with different causal direction.

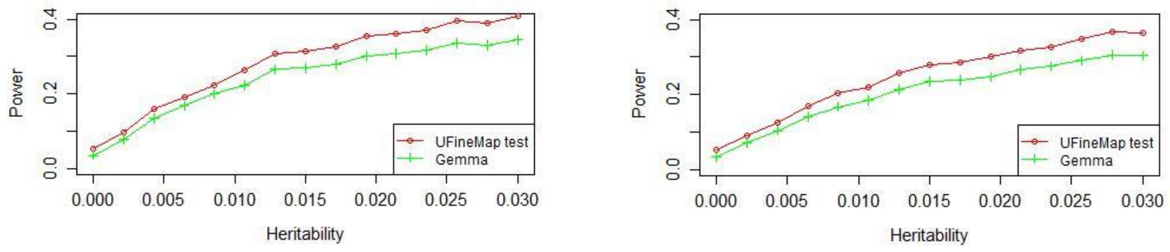

Figure 3.4. (a) Power comparison for uFineMap test with same causal direction.

(b) Power comparison for uFineMap test with different causal direction.

Figure 3.3 and Figure 3.4 present the power comparison of competing methods under same direction and different direction settings respectively. The similar patterns also occurred at a marker wise tests comparison. To make the presentation concise, we only show the result of regional tests, and the result of marker wise tests can be found in Figure

3.5 and 3.6. We can draw the conclusion that all three methods are robust with respect to causal variants direction. But our uHDSet test is almost uniformly more powerful than Gemma and famSKAT for SeqSIMLA simulation data.

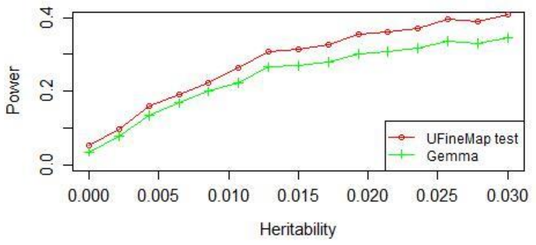

Figure 3.5. Power comparison for uFineMap test with same causal direction

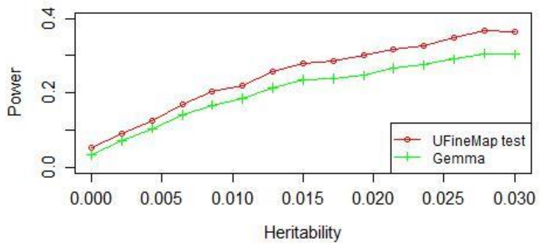

Figure 3.6. Power comparison for uFineMap test with different causal direction

We also provide power comparison for different sample size (Figure 3.7 and 3.8).

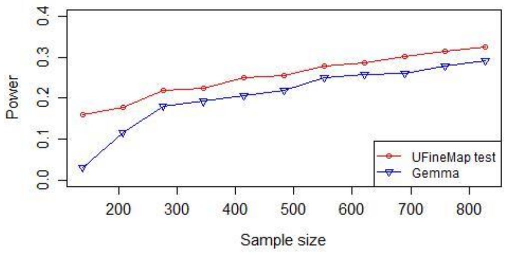

Figure 3.7. Power comparison for uFineMap test under different sample size

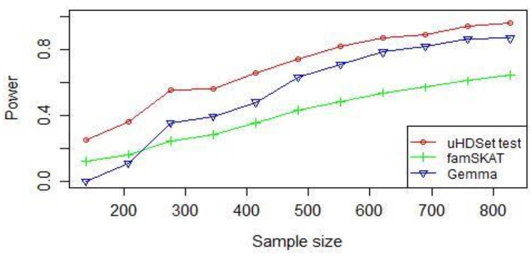

Figure 3.8. Power comparison for uHDSet test under different sample size

Table 3.1. The computational time for three selected methods

uHDSet test	famSKAT	Gemma
505s(1 core)/71.5s(8 cores)	1.04s	0.617s

Note the computational time is of one core is non-parallel computing. The testing data are generated by SeqSIMLA2 simulation with 552 samples and 1000 SNPs. CPU: Intel® Core™ i5-2410M @ 2.30GHz

3.6 Framingham heart study data

To demonstrate the effectiveness of our methods for real genetic variants detection, we applied them to the analysis of sequence data of Framingham Heart Study. This dataset contains both GWAS and next generation sequencing (NGS) data from 4229 subjects with HipBMD data. We used the FISH (Zhang, Pei et al. 2014) method for genotype imputation and selected HipBMD as the phenotype data. After quality control, we obtained 3322 individuals with 6,500,475 SNPs in total. We apply two kinds of data analysis strategies: whole genome analysis and pathway based analysis.

We present figure for PC1 vs PC2 (Figure 3.9) which stands for population structure.

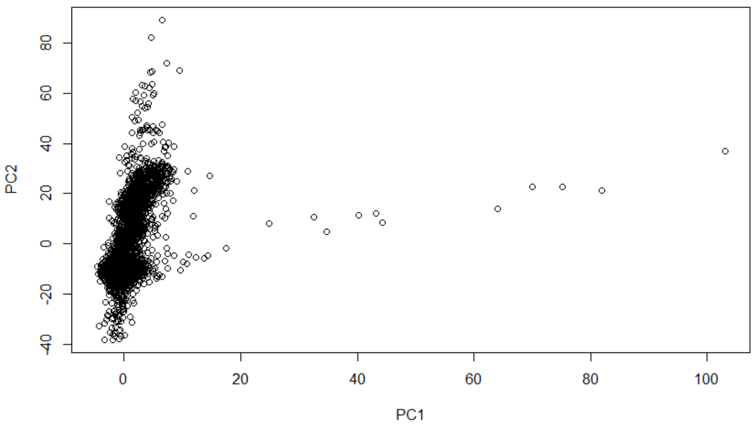

Figure 3.9. The PC1 vs PC2 plot for Framingham Heart Study genotype data

Figure 3.9 indicates that there is no significant population stratification for this data set. It is mainly due to the samples that are homogeneous.

3.6.1 Whole genome analysis

We separate each chromosome into several genetic windows and then apply our uFineMap and uHDSet tests in each window. We set the window size to be 100kb base pairs. After the separation, the whole genome is separated by a total number of 16514 sets of markers. The phenotype is adjusted by the covariates and the top 10 principle components of the genotype before the application of the proposed method. Following the same process as in the simulation studies, we obtain the results and draw the Manhattan plots for 22 chromosomes, as shown in Figure 3.10 and Figure 3.11 respectively.

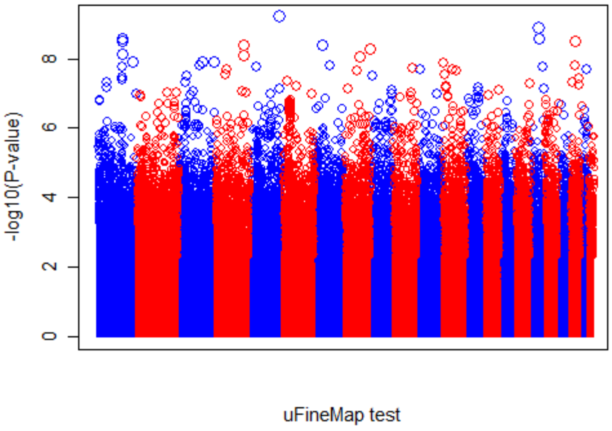

Figure 3.10. The Manhattan plot for uFineMap test of 22 chromosomes. Each point represents p-value of its corresponding SNP.

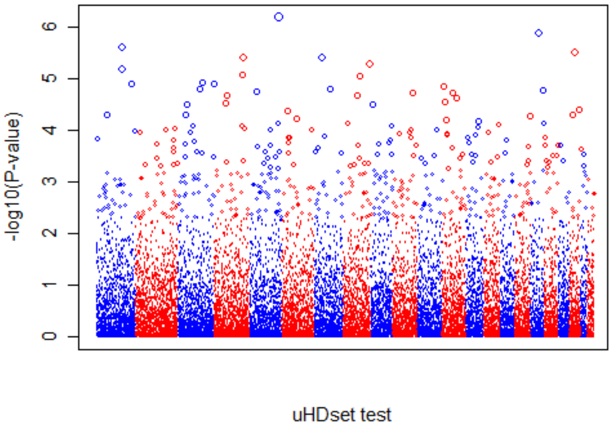

Figure 3.11. The Manhattan plot for uHDSet test of 22 chromosomes. Each point represents p-value of a 100kb window SNPs region.

By combining the overlapped region of Figure 3.10 and Figure 3.11, the uHDSet test report 68 regions of highest susceptibility that exceed a p-value threshold of 0.001. The reported p-value is based on the whole regional test. According to GeneCards websites, there are 11 genes (Table 3.2) within the selected regions that are associated with BMD or osteoporosis disease, which further confirms our findings. However, these 11 genes are missed by the use of famSKAT and Gemma method. The reported p-value of Gemma is generated by the minimal p-value after Bonferroni correction for the SNPs within the region.

Table 3.2. The selected susceptibility genes by uHDSet test

Gene	Chromosome	uHDSet p-value	famSKAT p-value	Gemma p-value
DNM3	1	2.47E-06	0.071107033	0.963871
APOB	2	7.43E-05	0.018075521	0.044156
ERC1	12	0.000154572	0.075876014	0.54554
SRD5A1	5	0.000267385	0.227392554	1
NR3C2	4	0.000317415	0.884812719	0.287339
PLCG1	20	0.000487724	0.022591921	1
INSIG2	2	0.00067805	0.73450689	0.29285
CYP24A1	20	0.000719511	0.132626874	1
ITGA1	5	0.000794757	0.143515502	1
BMPR2	2	0.000901023	0.762703102	0.729078
WNT4	1	0.000940191	0.602006435	0.718623

For the marker wise test, the uFineMap test report 82 susceptible SNPs that exceed a p-value threshold of 10^{-5}. Table 3.3 shows the 6 reported SNPs that are associated with BMD or osteoporosis disease according to GeneCards websites.

Table 3.3. The selected susceptibility SNPs by uFineMap test

SNPs	Gene	Chromosome	uFineMap	Gemma

rs11571334	ALOX12	17	4.47E-07	4.68E-05
rs3136452	F2	11	5.39E-07	8.37E-05
rs1264891	OVGP1	1	2.36E-06	5.53E-05
rs10513003	ITGA1	5	4.38E-06	2.99E-05
rs1491717	GC	4	5.17E-06	7.43E-05
rs235766	BMP2	20	5.67E-06	2.99E-05

3.6.2 Pathway analysis

To further illustrate the benefit of the uHDSet test, we collect 880 pathways from KEGG, REACTOME and BIOCARTA pathway analysis databases. We first extract genes belonging to each pathway, then select the corresponding SNPs. The selected SNPs of a specific pathway are combined to form the design matrix for association tests. We list 6 most significant pathways that pass p-value cut-off 10^{-3} in Table 3.4 for which the prominent famSKAT methods fails to detect. The two P38/MAPK pathways were previously found to play a critical role by other publications (Lee, Suh et al. 2008, Kim, Kim et al. 2013). Endogenous Sterols pathway is also related with BMD reported by another study (Warriner and Saag 2013). Chemokines pathway is important regulator in development, homeostasis and pathophysiological processes associated with osteoporosis (Lazennec and Richmond 2010).

Table 3.4. The selected functional pathways by uHDSet test only

Pathway name	uHDSet p-value	famSKT p-value
REACTOME_FACILITATIVE_NA_INDEPENDENT_GLUCOSE_TRANSPORTERS	5.00E-05	0.05809
REACTOME_ACTIVATED_TAK1_MEDIATES_P38_MAPK_ACTIVATION	7.00E-05	0.05635

REACTOME_P38MAPK_EVENTS	8.00E-05	0.09401
REACTOME_ENDOGENOUS_STEROLS	0.00016	0.00110
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	3.00E-04	0.07827
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES	0.00065	0.13751

Each p-value in Table 3.4 is generated based on a whole pathway-based region. It can be seen that, our uHDSet method is more powerful than famSKAT in identifying significant pathways which contain a relatively large number of genetic markers.

3.7 Conclusion

Some promising association tests with the adjustment of family structure have been established on the LDSs (low dimensional sets). However, these methods would suffer power loss in high dimensional data. To overcome the limitations of these tests, we propose the uFineMap and uHDSet test for assessing the significance of the HDSs with cryptic relatedness, which are based on novel scaled linear mixed sparse regression. The proposed tests are designed to address the challenge of variants detection under complex pedigree structures, which implement an explicit way to appropriately control the Type I error rate at both single marker level and SNPs set level.

The promising results of testing on both simulated and real data indicate that the uFineMap and uHDSet test yield a considerably higher statistical power gain in comparison to other competing methods, especially for high dimensional data with cryptic relatedness. The uFineMap test can pinpoint single susceptible variants with higher resolutions, even for rare functional variants. In addition, our methods also maintain substantial power for detecting susceptibility variants in low dimensional data or large samples. Last but not least, our methods can identify both rare and common variants efficiently.

One limitation of the proposed method is that we assume a linear mixed relationship between phenotype and genotype, which might not be true in the real world. Therefore, nonlinear regression models with adjustment of relatedness and population stratification may be more suitable. In addition, the overall computational complexity is $\mathrm{O}(n^2m^3)$, which is much longer than simply solving the sparse linear mixed model or other efficient methods designed for LDSs, particularly for extremely large data. To compensate for this issue, parallel computing is implemented to reduce the total computational time for large scale genetic data analysis.

CHAPTER 4 GENERALIZED UNIFIED SPARSE REGRESSION

4.1 Introduction

In statistics, a generalized linear mixed model (GLMM) is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. It extends the idea of linear mixed models to non-normal data, which is typical in clinical data applications. In this dissertation, we mainly focus on binary phenotype data, since this kind of data is popular in genetic research.

Mixed model is a statistical model containing both fixed effects and random effects. It is particular useful where measurements are made on clusters of related statistical units (family or longitudinal study). Assuming that the following linear model holds, maximizing the joint density for $\boldsymbol{\beta}$ and $\mathbf{u}$, gives Henderson's "mixed model equations" (MME) (Robinson 1991).

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$ and $u \sim N(\mathbf{0}, \mathbf{G})$

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} - \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

The solutions to the MME $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the best linear unbiased estimates (BLUE) and predictors (BLUP) for $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. This is a consequence of the Gauss-Markov theorem when the conditional variance of the outcome is not scalable to the identity matrix. When the conditional variance is known, then the inverse variance weighted least square estimate is BLUE. However, the conditional variance is rarely, if ever, known. So it is desirable to jointly estimate the variance and weighted parameter estimates when solving MME.

Another method used to fit such mixed models is that of the EM algorithm (Lindstrom and Bates 1988), where the variance components are treated as unobserved nuisance parameters in the joint likelihood.

Generalized linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, GLMMs is treated as an extension of generalized linear models (e.g., logistic regression) to include both fixed and random effects (hence mixed models).

Standard GLMM can be written by the following form

$$g(E[\mathbf{y} \mid \mathbf{u}]) = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{4.1}$$

where g is the link function for linear predictor, $y_i \mid \mathbf{u} \sim f_{Y_i|\mathbf{u}}(y_i \mid \mathbf{u})$

$$f_{Y_i|\mathbf{u}}(y_i \mid \mathbf{u}) = \exp\{[y_i \gamma_i - b(\gamma_i)] / \tau^2 - c(y_i, \tau)\}$$

However, fitting GLMMs via maximum likelihood involves integrating over the random effects. In general, those integrals cannot be expressed in analytical form. Various approximate methods have been developed, such as Penalized Quasi-Likelihood (PQL) (McCulloch 1997), Laplacian approximation and adaptive Gauss-Hermite quadrature (Bates, Maechler et al. 2014). In 2014, two papers (Groll and Tutz 2014, Schelldorfer, Meier et al. 2014) proposed a similar GLMM using L_1 norm penalty for high-dimensional variable selection.

In this article, we propose methods for high-dimensional GLMMs with group penalized regularization and L_p norm penalty. The L_p norm regularized GLMM is based on the previous USR methods; the group penalty norms are a group lasso type and sparse group lasso regularization. It can be solved with a cyclic coordinate descent optimization. For likelihood approximation of the GLMM, we implement Penalized Quasi-Likelihood (PQL) method which is efficient and robust.

To the best of our knowledge, there does not exist any literature devoted to develop structure sparse regularization model (e.g., group L1 norm and sparse group L1 norm) for high-dimensional GLMM. One of the main contribution of the dissertation is the generalization of USR model to incorporate non-Gaussian phenotype using GLMM; another one is that we provide a new toolkit, considering more general penalties with group lasso and sparse group lasso. It yields solutions that are sparse at both the group and individual feature levels in fitting GLMM. The structure sparse regularized GLMM can be naturally extended to gene based or pathway based association analysis.

4.2 Generalized linear mixed model with sparse regularizations

4.2.1 Objective function approximation

We assume that the non-Gaussian phenotype follows exponential family distribution, so the equation (4.1) holds. It is straightforward to write down the likelihood:

$$L = \int \prod_i f_{Y_i|\mathbf{u}}(y_i \mid \mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) d\mathbf{u} \tag{4.2}$$

where $\mathbf{u} \sim N(0, \mathbf{\Sigma})$

In (4.2), the integration is over the m-dimensional distribution of random effect $\mathbf{u}$. In general, it cannot be simplified further or evaluated in closed form. We adapt penalized quasi-likelihood (PQL) method to approximate and find the estimator of likelihood (McCulloch and Neuhaus 2001). The penalized quasi-likelihood (PQL) approach is the most common estimation procedure for the generalized linear mixed model (GLMM) (Jang, 2006). Central to PQL is the use of a Laplace approximation for evaluating the high-dimensional integral in the likelihood.

According to (4.2), we utilize Laplace approximation to approximate the log-likelihood of the GLMM

$$\begin{aligned} l(\boldsymbol{\beta},\mathbf{u},\mathbf{y}) &= \log L = \log\int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u}) f_U(\mathbf{u})d\mathbf{u} \\ &= \log\int \exp\{\log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u}) + \log f_U(\mathbf{u})\}d\mathbf{u} = \log\int\exp\{h(\mathbf{u})\}d\mathbf{u} \end{aligned} \tag{4.3}$$

With $h(\mathbf{u}) = \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u}) + \log f_U(\mathbf{u})$

The basic form of Laplace's approximation is based on a second-order Tylor series expansion:

$$\log\int\exp\{h(\mathbf{u})\}d\mathbf{u} \approx h(\mathbf{u}_0) + \frac{m}{2}\log(2\pi) - \frac{1}{2}\log\left| -\frac{\partial^2 h(\mathbf{u})}{\partial\mathbf{u}\partial\mathbf{u}^T}\right|, \text{ where } \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}}\Big|_{\mathbf{u}=\mathbf{u}_0} = \mathbf{0} \tag{4.4}$$

Combine (4.3) and (4.4), if we assume that $\mathbf{u} \sim N(0,\boldsymbol{\Sigma})$, then we have

$$\log f_U(\mathbf{u}) = -\frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}^{-1}\mathbf{u} - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}|, \text{ and}$$

$$h(\mathbf{u}) = \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u}) - \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}^{-1}\mathbf{u} - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}|$$

$$l(\mathbf{u}\,|\,\boldsymbol{\beta},\mathbf{X},\mathbf{y}) \approx \log(f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u})) - \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}^{-1}\mathbf{u} + C \tag{4.5}$$

Consider maximum likelihood of model (4.5), with additional regularization term. The general full model of GLMM with sparse regularization can be written as model (4.6).

$$\hat{\boldsymbol{\beta}} = \underset{\hat{\boldsymbol{\beta}}\in R^m}{\arg\min}\{-l(\boldsymbol{\beta},\mathbf{u},\mathbf{X},\mathbf{y}) + \lambda P(\boldsymbol{\beta})\} \tag{4.6}$$

Model (4.6) is the general form of sparse regularized GLMM.

4.2.2 Algorithms for L_p norm regularized GLMM

Given specific regularization term $P(\boldsymbol{\beta})$, we derive different algorithms to solve them. Since, L_1 norm penalty is already solved by (Groll and Tutz 2014, Schelldorfer, Meier et al. 2014), we consider the L_p (0<p<1) norm penalty. In another word, it is a generalization of USR model (Cao, Qin et al. 2014).

$$\hat{\boldsymbol{\beta}} = \underset{\hat{\boldsymbol{\beta}} \in R^m}{\arg\min}\{-l(\boldsymbol{\beta},\mathbf{u},\mathbf{X},\mathbf{y}) + \lambda \|\boldsymbol{\beta}\|_p^p\} \tag{4.7}$$

To solve this model, we first investigate the likelihood of GLMM. Differentiating $l(\boldsymbol{\beta},\mathbf{u},\mathbf{X},\mathbf{y})$ in (4.5) with respect to $\mathbf{u}$ gives

$$\begin{aligned}\frac{\partial l}{\partial \mathbf{u}} &= \frac{\partial \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u})}{\partial \mathbf{u}} - \boldsymbol{\Sigma}^{-1}\mathbf{u} \\ &= \frac{1}{\tau^2}\mathbf{Z}^T\mathbf{M}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) - \boldsymbol{\Sigma}^{-1}\mathbf{u}\end{aligned} \tag{4.8}$$

where $\mathbf{M} = diag[v(\mu_i)g(\mu_i)^2]^{-1}$ and $\boldsymbol{\Delta} = diag[g(\mu_i)]$ with

$\mu_i = E[y_i\,|\,\mathbf{u}]$, $g(\mu_i)$ is the link function and $v(\mu_i)$ is the first derivative of the link function $g(\mu_i)$. For some models (e.g., binomial or Poisson), $\mathbf{M}\boldsymbol{\Delta} = \mathbf{I}$

Consequently, the optimal $\mathbf{u}$ should satisfy:

$$\frac{1}{\tau^2}\mathbf{Z}^T\mathbf{M}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}\mathbf{u} \tag{4.9}$$

Differentiating $l(\boldsymbol{\beta},\mathbf{u},\mathbf{X},\mathbf{y})$ in (4.5) with respect to $\boldsymbol{\beta}$ gives

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{\partial \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}\,|\,\mathbf{u}_0)}{\partial \boldsymbol{\beta}} + \frac{\partial}{\partial \boldsymbol{\beta}}\frac{1}{2}\log|\mathbf{Z}^T\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma}/\tau^2 + \mathbf{I}| \\ &= \frac{1}{\tau^2}\mathbf{X}^T\mathbf{M}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})\end{aligned} \tag{4.10}$$

Based on (4.9) and (4.10), we propose the generalized USR as follows.

The Algorithm for GLMM with L_p norm regularization (GLMM-Lp)

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…$m$

Step 2: The random effect $\mathbf{u}$ can be estimated as $\hat{\mathbf{u}}^{(r+1)} = \mathbf{\Sigma Z}^T \mathbf{M\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Step 3: Update $\hat{\boldsymbol{\alpha}}^{(r+1)} = \hat{\boldsymbol{\alpha}}^{(r)} - (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{M\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r+1)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Update regression coefficients $\hat{\boldsymbol{\beta}}^{(r+1)} = R_{\lambda\mu,1/2}(\hat{\boldsymbol{\beta}}^{(r)} - \frac{1}{2}\mathbf{X}^T\mathbf{M\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r+1)}, \hat{\boldsymbol{\alpha}}^{(r+1)})))$

Step 4: Apply the lower bounds to regularize $\hat{\boldsymbol{\beta}}^{(r+1)}$ and use the SCG algorithm (Zhang *et al.*, 2009) with the initial point $\hat{\boldsymbol{\beta}}^{(r+1)}$ to find the minimizer $\boldsymbol{\beta}_p^{(r+1)}$ of objective function (8)

Step 5: If $\| \boldsymbol{\beta}_p^{(r+1)} - \boldsymbol{\beta}_p^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step 2

In addition, we also modify the glmmLasso by replacing L_1 norm with the elastic net penalty. So the number of selected variables could be larger than the number of sample size. The GLMM-EN (GLMM with Elastic Net regularization) can be written as:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\hat{\boldsymbol{\beta}} \in R^m}{\arg\min}\{-l(\boldsymbol{\beta}, \mathbf{u}, \mathbf{X}, \mathbf{y}) + \lambda(1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \lambda\alpha\|\boldsymbol{\beta}\|_1\} \\ &= \underset{\hat{\boldsymbol{\beta}} \in R^m}{\arg\min}\{g(\boldsymbol{\beta}, \mathbf{u}, \mathbf{X}, \mathbf{y}) + \lambda\alpha\|\boldsymbol{\beta}\|_1\} \end{aligned} \tag{4.11}$$

We apply proximal gradient descent algorithm (Chen, Lin et al. 2012) to solve GLMM-EN, which is similar to soft-thresholding algorithm for lasso.

The Algorithm for GLMM-EN

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…$m$

Step 2: The random effect $\mathbf{u}$ can be estimated as $\hat{\mathbf{u}}^{(r+1)} = \mathbf{\Sigma Z}^T \mathbf{M\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Step 3: Update $\hat{\boldsymbol{\alpha}}^{(r+1)} = \hat{\boldsymbol{\alpha}}^{(r)} - (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{M\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r+1)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Step 4: Apply line search to get the optimal step size t

Start from $t = t_0$, repeat $t = \eta t$ ($0 < \eta < 1$), until satisfy:

$$g(\boldsymbol{\beta}^{(r)} - tG_t(\boldsymbol{\beta}^{(r)})) \le g(\boldsymbol{\beta}^{(r)}) - t\nabla g(\boldsymbol{\beta}^{(r)})^T G_t(\boldsymbol{\beta}^{(r)}) + \frac{t}{2} \| G_t(\boldsymbol{\beta}^{(r)}) \|_2^2$$

Step5: Update regression coefficients $\boldsymbol{\beta}^{(r+1)} = prox_t(\hat{\boldsymbol{\beta}}^{(r)} - t\nabla g(\boldsymbol{\beta}^{(r)}))$

where $prox_t(x) = \begin{cases} x - t & x > t \\ 0 & |x| < t \\ -x + t & x < -t \end{cases}$

Step 5: If $\| \boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step 2

4.2.3 Algorithms for group lasso regularized GLMM

Furthermore, to incorporate group wise sparsity and select groups of SNPs from genes, we propose group L_1 norm regularized GLMM (GLMM-GL). Suppose we can divide predictors into L groups, with p_g predictors in group g. The GLMM-GL can be written as.

$$\hat{\boldsymbol{\beta}} = \underset{\hat{\boldsymbol{\beta}} \in R^m}{\arg\min} \{ -l(\boldsymbol{\beta}, \mathbf{u}, \mathbf{X}, \mathbf{y}) + \lambda \sum_{g=1}^{L} \sqrt{p_g} \| \boldsymbol{\beta}_g \|_2 \} \tag{4.12}$$

To solve (4.12), we implement the idea of group-wise coordinate descent algorithm (Meier, Van De Geer et al. 2008, Friedman, Hastie et al. 2010) and propose the following algorithm for GLMM-GL.

The Algorithm for GLMM-GL

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…$m$

Step 2: The random effect $\mathbf{u}$ can be estimated as $\hat{\mathbf{u}}^{(r+1)} = \boldsymbol{\Sigma}\mathbf{Z}^T\mathbf{M}\boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Update $\hat{\boldsymbol{\alpha}}^{(r+1)} = \hat{\boldsymbol{\alpha}}^{(r)} - (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{M}\boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r+1)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Step 3: For *g* from 1 to *L*, update group-wise regression coefficients

Check if $\| \frac{\partial l}{\partial \hat{\boldsymbol{\beta}}_g^{(r)}} \|_2^2 \le \lambda \sqrt{p_g}$, g=g+1; otherwise continue to Step 4

Step 4: Define $$\mathbf{d}_g = \begin{cases} -\hat{\boldsymbol{\beta}}_g^{(r)} & \| \frac{\partial l}{\partial \hat{\boldsymbol{\beta}}_g^{(r)}} - h_g \hat{\boldsymbol{\beta}}_g^{(r)} \|_2 \le \lambda \sqrt{p_g} \\ -\frac{1}{h_g} \{ \frac{\partial l}{\partial \hat{\boldsymbol{\beta}}_g^{(r)}} - \lambda \sqrt{p_g} \frac{\frac{\partial l}{\partial \hat{\boldsymbol{\beta}}_g^{(r)}} - h_g \hat{\boldsymbol{\beta}}_g^{(r)}}{\| \frac{\partial l}{\partial \hat{\boldsymbol{\beta}}_g^{(r)}} - h_g \hat{\boldsymbol{\beta}}_g^{(r)} \|_2} \} & \mathit{otherwise} \end{cases}$$

Find t s.t., $S_\lambda(\hat{\boldsymbol{\beta}}_g^{(r)} + t\mathbf{d}_g) - S_\lambda(\hat{\boldsymbol{\beta}}_g^{(r)}) \le t\sigma\Delta$ using line search

Step 5: If $\| \boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step 2

4.2.4 Algorithms for sparse group lasso regularized GLMM

One major disadvantage of the group lasso penalty is that it does not yield sparsity within a group. That is, if a group of parameters is non-zero, they will all be non-zero. In this case, we need to provide additional penalty that can select predictors at both group and single marker level. Sparse group lasso penalty (Simon, Friedman et al. 2013) is widely used to overcome this issue of group lasso. The GLMM with sparse group lasso regularization (GLMM-SGL) can be expressed as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\hat{\boldsymbol{\beta}} \in R^m}{\arg\min} \{ -l(\boldsymbol{\beta}, \mathbf{u}, \mathbf{X}, \mathbf{y}) + \lambda(1-\alpha) \sum_{g=1}^{L} \sqrt{p_g} \| \boldsymbol{\beta}_g \|_2 + \lambda\alpha \| \boldsymbol{\beta} \|_1 \} \tag{4.13}$$

To solve (4.13), we implement the modified version of group-wise coordinate descent algorithm for sparse group lasso (Vincent and Hansen 2014). The detailed algorithm is shown below:

The Algorithm for GLMM-SGL

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…m

Step 2: The random effect $\mathbf{u}$ can be estimated as $\hat{\mathbf{u}}^{(r+1)} = \mathbf{\Sigma} \mathbf{Z}^T \mathbf{M} \mathbf{\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Update $\hat{\boldsymbol{\alpha}}^{(r+1)} = \hat{\boldsymbol{\alpha}}^{(r)} - (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \mathbf{M} \mathbf{\Delta}(\mathbf{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\mathbf{u}}^{(r+1)}, \hat{\boldsymbol{\alpha}}^{(r)}))$

Step 3: Outloop:

For J from 1 to L, update group-wise regression coefficients

Check if $\sqrt{k(\lambda\alpha, \mathbf{g}_J^{(r)})} \le \lambda(1-\alpha)\sqrt{p_J}$, let $\hat{\boldsymbol{\beta}}_J^{(r+1)} = \mathbf{0}$; otherwise continue to Step 4

where $\mathbf{g}_J^{(r)} = -\nabla l(\boldsymbol{\beta}_J^{(r)}, \mathbf{u}, \mathbf{X}_J, \mathbf{y})$ is the first derivative over the group J

Step 4: Innerloop:

Define $Q^{(J)}(\boldsymbol{\beta}_J^{(r)}) = \mathbf{X}_J^T \mathbf{g}_J^{(r)} + \frac{1}{2} \mathbf{X}_J^T \mathbf{H}_{JJ} \mathbf{X}_J$, where $\mathbf{H}_{JJ}$ is the submatrix of Hessian matrix, the index of $\mathbf{H}_{JJ}$ corresponds to index of group J

$\Phi^{(J)}(\boldsymbol{\beta}_J^{(r)}) = (1-\alpha)\sqrt{p_J} \, \| \boldsymbol{\beta}_J^{(r)} \|_2 + \alpha \| \boldsymbol{\beta}_J^{(r)} \|_1$

Let $\boldsymbol{\beta}_J^{inner} = \underset{\boldsymbol{\beta}_J^{inner} \in R^{nJ}}{\arg\min} \{ Q^{(J)}(\boldsymbol{\beta}_J^{(r)}) + \lambda \Phi^{(J)}(\boldsymbol{\beta}_J^{(r)}) \}$

Step 5: if $\boldsymbol{\beta}_J^{inner}$ satisfy inner loop criteria, proceed to Step 6; otherwise return to Step 4

Step 6: If $\| \boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)} \|_2 < \varepsilon.outloop$ stop; otherwise return to Step 2

4.2.5 Logistic mixed model with sparse regularizations

Binary phenotype or case-control study is one of the most widely used design in genomic data. Without loss of much generality, we narrow down our GLMM to be logistic mixed model in the following discussion.

Consider the odd ratio inference problem. In this case, the phenotype $y_i \in \{0,1\}$ stands for control($y_i = 0$) and case($y_i = 1$) respectively. The conditional probability of the phenotype is: $\Pr(\mathrm{y_i} = 0 \mid \mathbf{x}_i, \mathbf{z}_i) = \dfrac{1}{1+\exp(-\mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{z}_i^T \mathbf{u})}$,

$$\Pr(\mathrm{y_i} = 1 \mid \mathbf{x}_i, \mathbf{z}_i) = \frac{1}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u})}$$

To be explicit, the logistic mixed model is:

$$\mathrm{logit}(E(\mathbf{Y} \mid \mathbf{u})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\text{, where } \mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

To estimate the parameters, consider the integrated likelihood:

$$L(\boldsymbol{\beta}, \mathbf{u}) = \frac{1}{\sqrt{(2\pi)^m \,|\boldsymbol{\Sigma}|}} \int \exp[\sum_{i=1}^{n} (y_i \log(\frac{p_i(\boldsymbol{\beta}, \mathbf{u})}{1-p_i(\boldsymbol{\beta}, \mathbf{u})}) + \log(1-p_i(\boldsymbol{\beta}, \mathbf{u}))) - \frac{1}{2}\mathbf{u}^T \boldsymbol{\Sigma}^{-1}\mathbf{u}]d\mathbf{u}$$

where $p_i(\boldsymbol{\beta}, \mathbf{u}) = \dfrac{1}{1+\exp(-\mathbf{X}_i^T \boldsymbol{\beta} - \mathbf{Z}_i^T \mathbf{u})}$

Since the integral cannot be evaluated as a closed form, an alternative is to use a Laplace approximation. The corresponding log-likelihood can be expressed as:

$$l(\boldsymbol{\beta}) = \max_{\mathbf{u}}\{\sum_{i=1}^{n} [y_i \log(\frac{p_i(\boldsymbol{\beta}, \mathbf{u})}{1-p_i(\boldsymbol{\beta}, \mathbf{u})}) + \log(1-p_i(\boldsymbol{\beta}, \mathbf{u}))] - \frac{1}{2}\mathbf{u}^T \boldsymbol{\Sigma}^{-1}\mathbf{u}\}$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{y}-\mathbf{p}), \quad \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X}^T \mathbf{W} \mathbf{X}$$

Consequently, the logistic linear mixed model with sparse regularization can be expressed as:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in R^m} -l(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}) \tag{4.14}$$

To further extend our model for different prior knowledge, we consider different regularization functions as discussed above:

L_p norm: $P(\boldsymbol{\beta}) = \| \boldsymbol{\beta} \|_p^p$

Elastic-net penalty: $P(\boldsymbol{\beta}) = (1-\alpha) \| \boldsymbol{\beta} \|_2^2 + \alpha \| \boldsymbol{\beta} \|_1$

Group lasso: $P(\boldsymbol{\beta}) = \sum_{g=1}^{L} \sqrt{p_g} \| \boldsymbol{\beta}_g \|_2$

Sparse group lasso: $P(\boldsymbol{\beta}) = (1-\alpha) \sum_{g=1}^{L} \sqrt{p_g} \| \boldsymbol{\beta}_g \|_2 + \alpha \| \boldsymbol{\beta} \|_1$

4.2.6 Tuning parameter selection

As we discussed in **Section 2.2.8**, tuning parameter selection is an open problem. All the four models we proposed in this section need an appropriate tuning parameter selection strategy. However, the fitting term of GLMM is different from the least square regression loss function. The previous hybrid AIC and stability selection method is too conservative in parameter selection for GLMM loss function. In this chapter, we applied a less conservative method for the selection of parameters. We implemented cross validation with one standard error rule (1 SE rule) (Friedman, Hastie et al. 2001). The 1 SE rule is defined as selecting the most parsimonious model whose error is no more than one standard error above the error of the best model.

4.3 Simulation: Nuclear families with binary phenotype

To demonstrate the effectiveness of the group penalties, we proposed three types of simulation strategy and compared their performance. The first is without group structure, which is using the exactly same genotype generation procedure. The second one divides

1000 SNPs into 100 groups (each group has 10 SNPs). We set the SNPs' correlation within the same group to be 0.4, and between groups to be 0.1. In addition, we randomly assign 6 causal groups, and each group has different number of causal SNPs. The third strategy is to maintain the same group structure as the second one, but assign each causal group only one causal SNP.

We compare our method with "lme4" (Bates 2007), "glmnet" (Friedman, Hastie et al. 2009), "glmmLasso" (Schelldorfer, Meier et al. 2014) and "SKAT" (Wu, Lee et al. 2011). "lme4" solves generalized linear mixed model without sparse regularization; "glmnet" can handle the logistic regression with Lasso regularization, but it does not consider mixed effect for pedigree structure. "glmmLasso" provide a variable selection approach for generalized linear mixed models by L_1-penalized estimation. "SKAT" perform a kernel-regression-based association test for SNPs set.

4.3.1 Non-group structure simulation

We use the same genotype sampling procedure as in Section **2.3.2**. For the binary phenotype simulation, we employed the following liability threshold model. The only difference is the setting of the additional prevalence parameter.

$$\mathrm{Pr}_i = b\mathbf{X}_i\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_i \quad ,(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4.15}$$

$$\mathrm{Pr}(Y_i = 1) = \Phi^{-1}(\mathrm{Pr}_i)$$

Using this simulation, we compare competitive methods under three different causal SNPs settings. The first one is the "Sparse Group" model, where we randomly assign 15 causal groups, and each groups contains only one causal SNPs. The second is the "Dense Group" model, where we only define 6 causal groups but more than 4 causal SNPs are assigned for each causal group. The third one is the “Mixture Group” model which is a combination of “Sparse Group” and “Dense Group”, including 3 dense causal groups and 6 sparse causal groups.

To illustrate the robustness of our methods regarding to different causal direction of SNPs (i.e., protective SNPs and deleterious SNPs), we provide two options of $\boldsymbol{\beta}_0$: one is setting all the causal effect to be same direction, i.e., $\beta_{0j}=1$, for all j belongs to predetermined causal SNPs; the other one randomly assigns half of causal SNPs to have negative effect. The comparison of ROC curves is shown in Figure 4.1 and Figure 4.2 below and Table 4.1 compares the area under curve (AUC). Each point of ROC is generated by 500 times of replication.

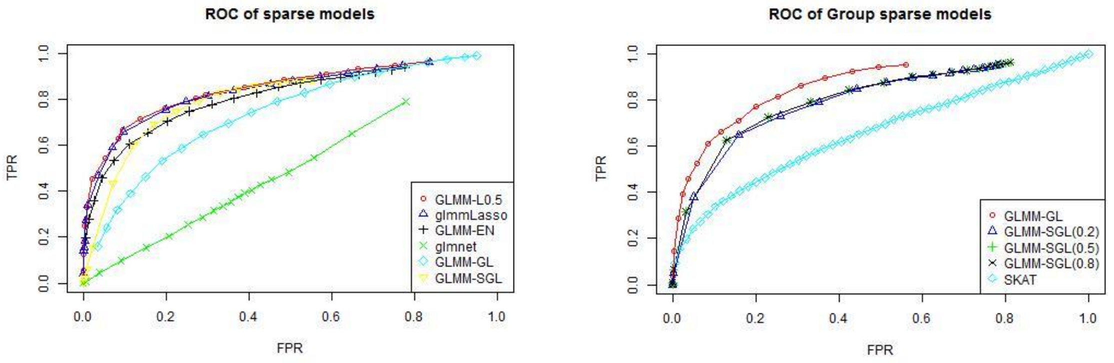

Figure 4.1. The ROC for non-group design, Heritability=0.5, “Sparse Group” design, same direction. **(a)** SNPs selection

(b) Groups of SNPs selection

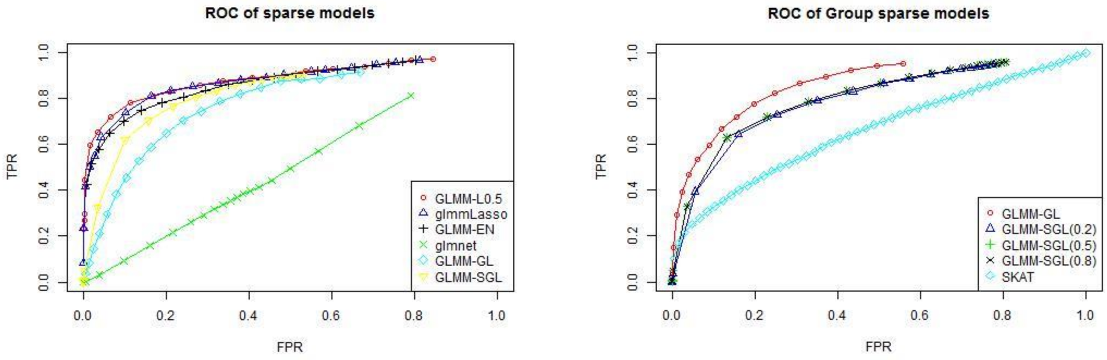

Figure 4.2. The ROC for non-group design, Heritability=0.5, “Sparse Group” design, different direction. **(a)** SNPs selection

(b) Groups of SNPs selection

In Figure 4.1(a), the TPR is defined by the number of selected true casual SNPs divided by the total number of true casual SNPs; and the FPR is defined by the number of selected false SNPs divided by the total number of false SNPs. On the other hand, in Figure 4.1(b), the TPR is defined by the number of selected true casual groups divided by the total number of true casual groups; and the FPR is defined by the number of selected false groups divided by the total number of false groups.

Figure 4.1(a) and Table 4.1 indicate that the GLMM-L0.5 is comparable to glmmLasso and GLMM-EN, and all the three GLMM sparse regularized methods outperform the glmnet without random effect. Under “Sparse Group”, in terms of single causal SNPs identification power, all the non-group sparse methods perform better than GLMM-GL and GLMM-SGL. GLMM-SGL, where the regularization parameter is set to be $\alpha = 0.5$, being more powerful than GLMM-GL.

Figure 4.1(b) and Table 4.2 show that the GLMM-GL yields higher power than GLMM-SGL in terms of group-wise discovery rate. However, the GLMM-GL method is less powerful under "Sparse Group" setting. We can also conclude that all the GLMM-SGL methods have similar detection power regarding to causal groups.

Table 4.1. The ROC for causal SNPs selection under non-group design

	GLMM-L0.5	glmmLasso	GLMM-EN	glmnet	GLMM-GL	GLMM-SGL
Sparse Group, Same Direction	0.8381	0.8350	0.8253	0.4028	0.7573	0.8261
Sparse Group, Different Direction	0.8336	0.8343	0.8229	0.4103	0.7617	0.8254
Dense Group, Same Direction	0.8085	0.8133	0.8058	0.6863	0.9213	0.9163
Dense Group, Different Direction	0.8234	0.8158	0.8175	0.6971	0.9351	0.9250

Table 4.2. The ROC for causal groups selection under non-group design

	GLMM-GL	GLMM-SGL(0.2)	GLMM-SGL(0.5)	GLMM-SGL(0.8)	SKAT
Sparse Group, Same Direction	0.8032	0.7902	0.798	0.798	0.6356
Sparse Group, Different Direction	0.8083	0.7856	0.7951	0.7951	0.638
Dense Group, Same Direction	0.9201	0.9587	0.9592	0.9592	0.8497
Dense Group, Different Direction	0.9356	0.962	0.964	0.964	0.8348

Figure 4.3 and Figure 4.4 show the ROC comparison under "Dense Group" design. The GLMM-GL clearly overwhelmed other methods in both single causal SNPs and causal groups discovery. GLMM-SGLs are just second to GLMM-GL, which is under our expectation. Since the causal SNPs forms a clear group structure, group regularized GLMM methods are favorable.

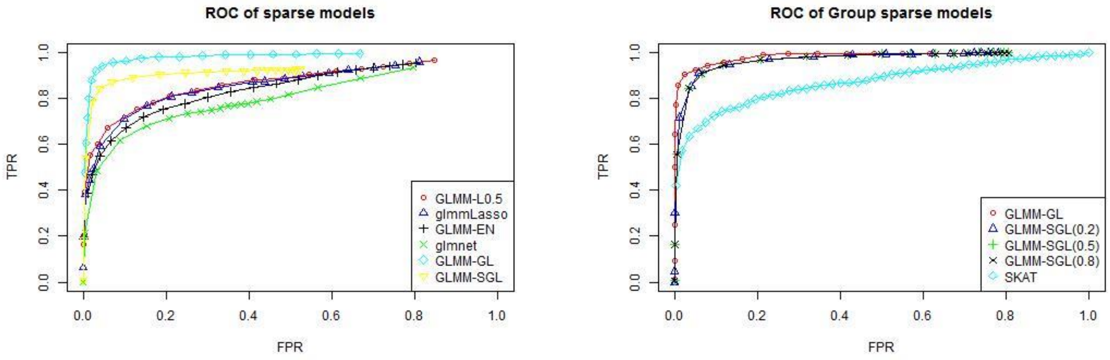

Figure 4.3. The ROC for non-group design, Heritability=0.5, “Dense Group” design, same direction. **(a)** SNPs selection

(b) Groups of SNPs selection

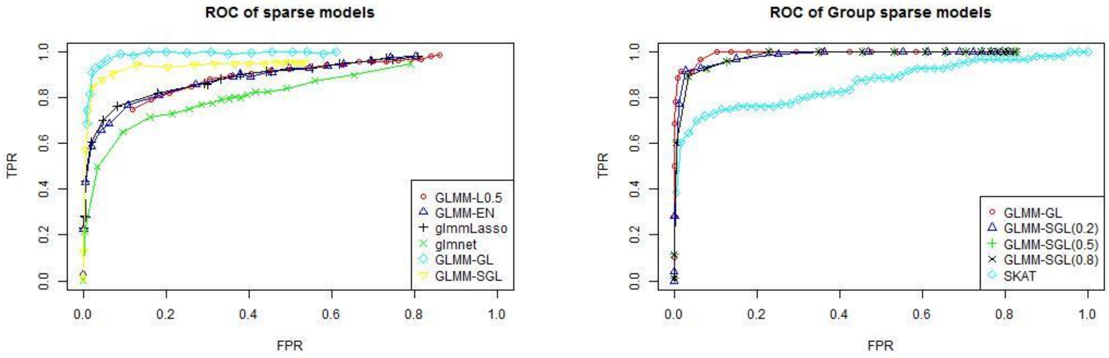

Figure 4.4. The ROC for non-group design, Heritability=0.5, “Dense Group” design, different direction. **(a)** SNPs selection

(b) Groups of SNPs selection

Theoretically, all the methods in comparison are not sensitive to causal direction of SNPs. It is proved by the closeness of performance between Figure 4.1 and Figure 4.2. For sake of simplicity, we only show the result of same causal direction setting in the remaining section.

4.3.2 Group structure simulation

To control the correlation between SNPs, we used multivariate Gaussian distribution to simulate underlying distribution of correlated SNPs. At each SNP, we chose the same minor allele frequency as in the YRI and CEU mixed haplotype data. In this simulated admixture, the frequencies of minor alleles range from 0.0011 to 0.5722. This simulation design includes three major steps.

Step1. Generate parental dataset

For each family, we generated father and mother independently. Each subject is composed of two haplotypes; For each haplotype, we assign a multi-variate Gaussian distribution $G_{n\times m} \sim N(0, \Gamma_{m\times m})$, where $\Gamma_{m\times m}$ is the predetermined covariate matrix of SNPs. We set within group correlation to be 0.4 and between group correlation to be 0.1 in $\Gamma_{m\times m}$. For each haplotype of SNP *j*, if $G_j > \Phi^{-1}(MAF_j)$, we assign the corresponding haplotype to be 1, otherwise we set the corresponding haplotype to be 0.

Step2. Generate nuclear family with two children

Two children are generated for each family. To generate one child, we randomly selected one haplotype from father and the other from mother. We simulated 150 families with the same family structure, which is composed of two parents with two children.

Step3. Generate trait values

To be explicit, for each person, we use the following model to generate trait values.

$$Y_i = b\mathbf{X}_i\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_i,\ (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (2.6)$$

where $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_N)$

In our simulation, the covariate matrix for each family is $\mathbf{\Sigma}_j = \frac{2}{3}\mathbf{\Phi} + \frac{1}{3}\mathbf{I}$

where $\mathbf{\Phi} = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$

is the kinship matrix. To simulate different heritability $H^2 = 0.5$ or 0.1, the true model we used is formula (4.15)

where $b = \dfrac{H}{\sqrt{(1-H^2)Var(\mathbf{X\beta})}}$

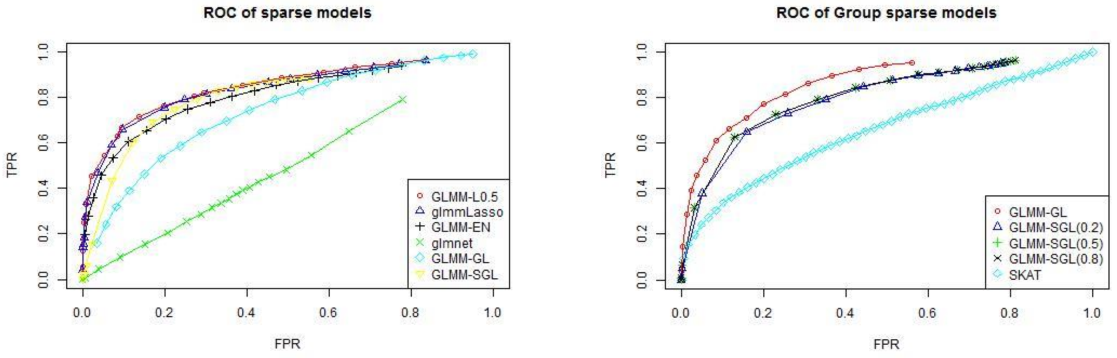

Figure 4.5. The ROC for group design, Heritability=0.5, “Sparse Group” design, same direction. **(a)** SNPs selection

(b) Groups of SNPs selection

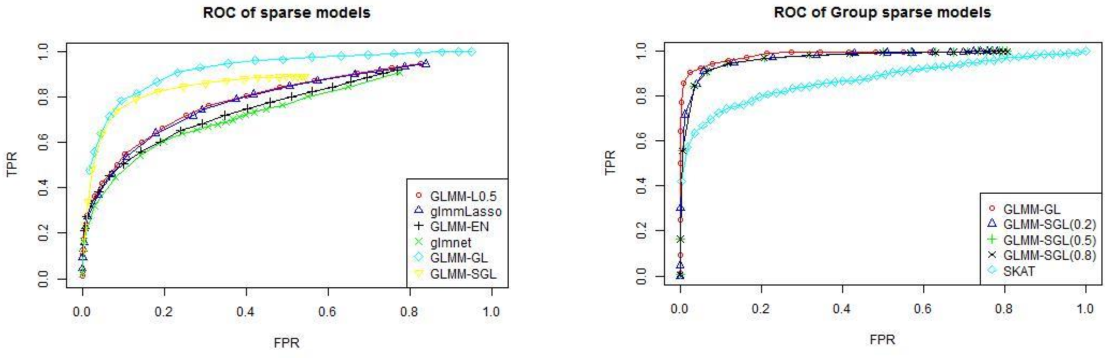

Figure 4.6. The ROC for group design, Heritability=0.5, “Dense Group” design, different direction. **(a)** SNPs selection

(b) Groups of SNPs selection

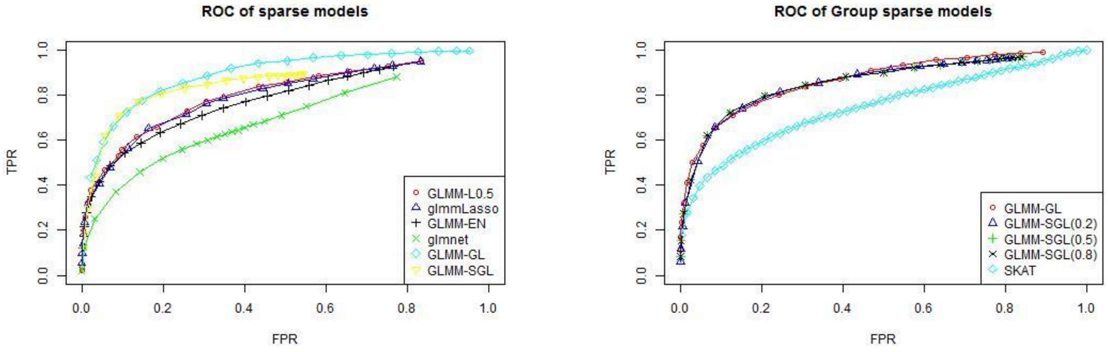

Figure 4.7. The ROC for group design, Heritability=0.5, “Mixture Group” design, different direction. **(a)** SNPs selection

(b) Groups of SNPs selection

Table 4.3. The AUC for causal SNPs selection under group design

	GLMM-L0.5	glmmLasso	GLMM-EN	glmnet	GLMM-GL	GLMM-SGL
Sparse Group	0.7800	0.7762	0.7668	0.4084	0.7034	0.8034
Dense Group	0.7308	0.7415	0.7305	0.6330	0.8837	0.8611
Mixtrue Group	0.7473	0.7336	0.7284	0.5848	0.8535	0.8481

Table 4.4. The AUC for causal groups selection under group design

	GLMM-GL	GLMM-SGL(0.2)	GLMM-SGL(0.5)	GLMM-SGL(0.8)	SKAT
Sparse Group	0.8186	0.7876	0.7958	0.7958	0.6660

Dense Group	0.9044	0.9020	0.9050	0.9050	0.7747
Mixtrue Group	0.8517	0.8465	0.8480	0.8480	0.7235

Results in Figure 4.5 to 4.7 illustrate the trend of ROC under three different simulation designs (“Sparse Group”, “Dense Group” and “Mixture Group”). The AUC (area under curve) in Table 4.3 and 4.4 indicate the similar pattern as in Section **4.3.1**. So far, one major conclusion is that the AUC and performance of GLMM-GL and GLMM-SGL mainly depend on the distribution of causal SNPs in each group. If causal SNPs are concentrated within a few groups, the causal groups’ detection power would be higher. In group-wise detection power, GLMM-SGL is comparable to GLMM-GL. GLMM-SGL is expected to be a robust method between single variable sparse regularized GLMM and group regularized GLMM. In addition, the GLMM-Lp is as good as glmmLasso, and has even better computational efficiency.

4.4 Framingham heart study data

To illustrate the effectiveness of the group regularized GLMM methods for real genetic variants identification, we applied them to the analysis of Framingham Heart Study data. This dataset contains sequence data of 8990 subjects which are generated by Affymetrix 500K genotyping platform.

The initial step in data preprocessing is to exclude individuals or SNPs with too much missing genotype data. We applied a filtering strategy which excludes individuals or SNPs with more than 10% missing genotypes. Then we used IMPUTE2 (Howie, Donnelly et al. 2009) for genotype imputation based on the filtered data. Consequently, we got 8915 subjects’ genotype data with 476907 annotated SNPs. Finally, we obtained 1519 individuals in total, 499 subjects with confirmed fracture and 1020 without fracture. We grouped SNPs by their corresponding genes and applied a whole genome analysis strategy to analyze each chromosome as a whole data. The tuning parameter is determined by 1 SE rule discussed in section **4.2.6**.

By using GLMM-SGL method, we totally selected 169 genes. According to GeneCards websites, there are 26 genes (Table 4.5) within the selected regions that are associated with BMD or osteoporosis disease. We selected 428 SNPs by GLMM-L0.5 method, and 31 of them are reported by literature, which is listed in Table 4.6.

Table 4.5. The selected susceptibility genes by GLMM-SGL

Selected Genes	P-value of SKAT
A4GALT	0.170945311
CTDP1	0.010612726
ITGB2	0.041121675
FMN1	0.19998
NIN	0.196937896
NR1H2	0.142696666
INSR	0.130728863
RUNX1	0.075256402
MC2R	0.101311204
CLDN14	0.04476686
SALL4	0.06442441
FOXG1	0.01359052
WRB	0.097248032
ANXA2	0.106435907
DIO2	0.036973428
ATP8B1	0.025232579
RNASEH2B	0.19998
MB	0.01302217
TTR	0.153804318
TMPRSS6	0.054912598
CXXC1	0.066942867
NKX2-1	0.007047425
PCNT	0.19998
GRAP2	0.189967487
MARK3	0.082637776

CACNA1A	0.0193284

Table 4.6. The selected susceptibility genes by GLMM-L0.5

rs number	Gene
rs7874142	COL5A1
rs10859155	DCN
rs14144	POLR1D
rs571118	KL
rs9525625	TNFSF11
rs1854521	SLC10A2
rs1335808	COL4A1
rs1955711	FBXO33
rs10873099	OTX2
rs941845	CATSPERB
rs3945958	CHGA
rs1551868	OCA2
rs409668	LIPC
rs11634686	SMAD3
rs17785209	THSD4
rs1630373	ERCC4
rs1861868	FTO
rs1493892	ADAMTS18
rs11867674	CA10
rs4622543	PRKCA
rs12452379	TIMP2
rs4277413	DCC
rs12959396	TNFRSF11A
rs9950037	CDH7
rs17827157	DOK6
rs7245376	CTDP1
rs6085696	BMP2
rs5011374	PLCB4
rs1888406	CXADR
rs2834694	RUNX1
rs6001491	PDGFB

In addition, we also performed KEGG pathway enrichment analysis and found 5 enriched pathways by 169 selected genes (Table 4.7). Gene expression level changes in neuroactive ligand-receptor interaction pathway is associated with osteoporosis (Liu, Zhu et al. 2015).

Table 4.7. KEGG pathway enrichment analysis

PathwayName	#Gene	Statistics
Neuroactive ligand-receptor interaction	5	rawP=0.0017;adjP=0.0204
Cell adhesion molecules (CAMs)	3	rawP=0.0086;adjP=0.0300
Axon guidance	3	rawP=0.0079;adjP=0.0300
Type II diabetes mellitus	2	rawP=0.0100;adjP=0.0300
RNA degradation	2	rawP=0.0210;adjP=0.0420
Long-term depression	2	rawP=0.0204;adjP=0.0420

4.5 Conclusion

Low-dimensional GLMMs and high-dimensional generalized linear model have been extensively studied in recent years. In addition, a few sparse regularized methods were proposed for high-dimensional GLMMs variable selection. However, all of the above methods ignore the effect of grouped variables, which are believed to have sparse effects both on a group and within group level.

We developed algorithms for solving the group lasso and sparse group lasso optimization problem with a GLMM loss function. In addition, we extend USR model to non-Gaussian phenotype with an efficient algorithm. The proposed methods can handle high-dimensional GLMMs and allow for correlated grouped predictors simultaneously. The promising results on both simulated and real data indicate that the GLMM-GL and GLMM-SGL yield higher power in grouped SNPs detection, especially for high dimensional data with cryptic relatedness.

However, there are still some open questions remain. Further theoretical research is needed to find ways of constructing significance test for high-dimensional predictors with

GLMM loss function. On the other hand, it deserves further investigation on how to choose the optimal tuning parameters and find the solution path of group lasso and sparse group lasso. Last but not least, allowing for overlapped group structure of SNPs within overlapped pathways is necessary for pathway-based approach.

CHAPTER 5 SUMMARY

5.1 Overview

In this thesis work, we are committed to developing a novel unified sparse regression (USR) approach for causal variants identification from family based sequencing data, especially under high dimensional settings.

For Specific Aim 1, we develop a novel L_p norm based sparse regression model (USR) for identifying genetic variants sets influencing on specific phenotype. USR is an effective method to incorporate prior information and jointly adjust for relatedness, population structure and environmental covariates. Our algorithm adopts a modified kinship matrix to account for the confounding of complex relationship between pedigree members on a quantitative trait (Thompson and Shaw 1990). For the data of cryptic relatedness, we infer the kinship matrix by the REAP algorithm (Thornton, Tang et al. 2012). Meanwhile, our USR models population structure and other environmental covariates as fixed effects. To allow proper sparsity and incorporate prior knowledge, our USR algorithm applies a weighted regularization with L_p norm ($0< p <1$) to select sparse representation – a sparse subset from a large number (>sample size) of markers. Our algorithm can automatically search for a sparse representation and allow users to determine the size of output set. As demonstrated by extensive sequence data analyses, our USR appears more effective than many existent sparse regression models.

For Specific Aim 2, we develop theories to assess statistical significance and control the Type I error rates of genetic variants and gene sets by using scaled Lp norm regularized sparse regression model. Besides the feature selection method of USR, we need to further evaluate the asymptotic property of USR estimators, in order to control the Type I error rate or Family-wise error rate. We first present a unified test (uFineMap) for accurately localizing causal loci. The uFineMap is a marker wise test under a scaled sparse linear mixed regression, which jointly models marker wise effect, relatedness and population stratification. It applies scaled L_p ($0 < p <1$) norm regularization to generate a de-biased solution. Next, we present a unified test (uHDSet) for identifying high-dimensional sparse associations in deep sequencing genomic data of related individuals. The uHDset integrates the marker wise statistics of the uFineMap to identify susceptible high-dimensional marker sets. In the uHDSet, the dependence among markers is modeled to appropriately control set-based Type I error rates. Under extensive simulations, the uFineMap outperformed the GEMMA (Zhou and Stephens 2012) and a Scaled Lasso based method (Javanmard and Montanari). The uHDSet yields higher statistical power than famSKAT and GEMMA.

For Specific Aim 3, we accommodate to non-Gaussian phenotype data and grouped high-dimensional covariates by developing generalized unified sparse regression model for association analysis. In a wide range of practical genetic data settings, binary phenotypes are commonly appeared other than continuous phenotype. It is critical to extend our USR methods to accommodate binary phenotypes. Although the non-Gaussian property and pedigree structure introduce analytical challenges for incorporating case-control data analysis into the regression model, there are several methods working on it. Generalized linear mixed models (GLMMs) are widely used to model correlated and clustered non-Gaussian responses (Jang and Lim 2006, Bates 2007). Since there is no analytical form of the original likelihood function of Generalized linear mixed model, we adopted the idea of PQL (Penalized quasi-likelihood) method (Mammen and van de Geer 1997) for the likelihood approximation with L_p norm regularization. Consequently, we proposed and generalized a new USR method to efficiently incorporate binary phenotype data while adjusting for high-dimensional grouped variants and random relatedness structure.

The general workflow for the three aims of this thesis is displayed in Figure 5.1.

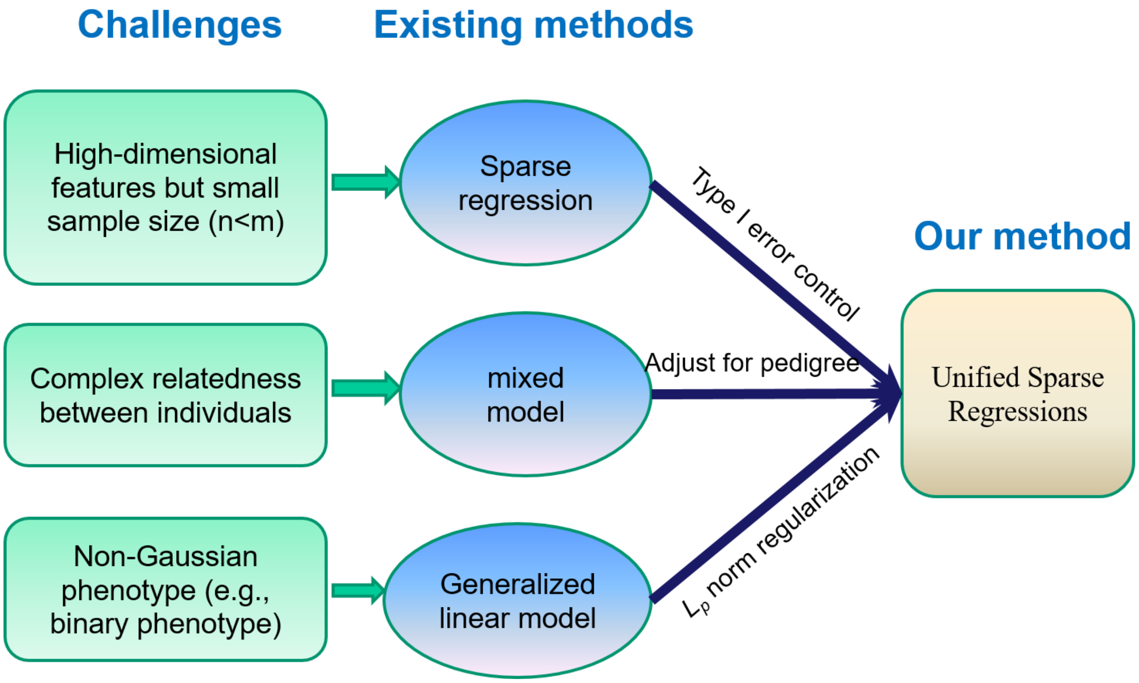
Challenges
Existing methods
High-dimensional features but small sample size (n<m)
Sparse regression
Complex relatedness between individuals
mixed model
Non-Gaussian phenotype (e.g., binary phenotype)
Generalized linear model
Type I error control
Adjust for pedigree
L_p norm regularization
Our method
Unified Sparse Regressions

Figure 5.1 Overview of USR models

5.2 Future work

Although we have developed sophisticated methods for genetic data analysis, there are still many open questions to be answered. The generalized linear mixed model with group sparse regularization requires a more efficient parameter selection strategy, in order to control the Type I error rate, family-wise error rate or false discovery rate. Further effort is needed to investigate the solution path of GLMM-GL and GLMM-SGL methods.

In addition to a single type of phenotypes, pleiotropic effect of high-dimensional genetic variants with random relatedness is another important extension. Pleiotropy occurs when one gene influences two or more seemingly unrelated phenotypic traits. Consequently, a method that can integrate related phenotypes together would be more powerful to detect the genetic mutation which has pleiotropic effect.

Last but not the least, to increase the computational efficiency of USR test method is also important since current computational complexity is $\mathrm{O}(n^2m^3)$, which is nearly impractical for large genetic region even with parallel computing. A better way to divide and group variants can dramatically reduce computational cost which is worth of further investigation.

REFERENCES

Amos, C. I. (1994). "Robust variance-components approach for assessing genetic linkage in pedigrees." American journal of human genetics **54**(3): 535.

Aulchenko, Y. S., et al. (2007). "GenABEL: an R library for genome-wide association analysis." Bioinformatics **23**(10): 1294-1296.

Ayers, K. L. and H. J. Cordell (2013). "Identification of grouped rare and common variants via penalized logistic regression." Genetic epidemiology **37**(6): 592-602.

Bates, D. (2007). "Linear mixed model implementation in lme4." Manuscript, University of Wisconsin-Madison.

Bates, D. (2014). "Computational methods for mixed models." LME4: Mixed-Effects Modeling with R: 99-118.

Bates, D., et al. (2014). "lme4: Linear mixed-effects models using Eigen and S4." R package version **1**(7).

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological): 289-300.

Benjamini, Y. and Y. Hochberg (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics." Journal of Educational and Behavioral Statistics **25**(1): 60-83.

Bühlmann, P. (2013). "Statistical significance in high-dimensional linear models." Bernoulli **19**(4): 1212-1242.

Bühlmann, P. and S. Van De Geer (2011). Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media.

Candes, E. J. and T. Tao (2005). "Decoding by linear programming." Information Theory, IEEE Transactions on **51**(12): 4203-4215.

Cao, S., et al. (2013). A generalized sparse regression model with adjustment of pedigree structure for variant detection from next generation sequencing data. Proceedings of the

International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. Wshington DC, USA, ACM**:** 191-200.

Cao, S., et al. (2014). "A unified sparse representation for sequence variant identification for complex traits." Genetic epidemiology **38**(8): 671-679.

Cao, S., et al. (2015). "Unified tests for fine-scale mapping and identifying sparse high-dimensional sequence associations." Bioinformatics: btv586.

Cetin, M. C. and A. Erar (2002). "Variable selection with Akaike information criteria: a comparative study." Hacettepe Journal of Mathematics and Statistics **31**: 89-97.

Chen, H., et al. (2013). "Sequence kernel association test for quantitative traits in family samples." Genetic epidemiology **37**(2): 196-204.

Chen, X. and H. Ishwaran (2012). "Random forests for genomic data analysis." Genomics **99**(6): 323-329.

Chen, X., et al. (2012). "Smoothing proximal gradient method for general structured sparse regression." The Annals of Applied Statistics **6**(2): 719-752.

Chen, X., et al. (2010). "Lower Bound Theory of Nonzero Entries in Solutions of \ell_2-\ell_p Minimization." SIAM Journal on Scientific Computing **32**(5): 2832-2852.

Chung, R. H., et al. (2015). "SeqSIMLA2: Simulating Correlated Quantitative Traits Accounting for Shared Environmental Effects in User ‐ Specified Pedigree Structure." Genetic epidemiology **39**(1): 20-24.

Devlin, B. and K. Roeder (1999). "Genomic control for association studies." Biometrics **55**(4): 997-1004.

Dong, C., et al. (2012). "Follow-up association study of linkage regions reveals multiple candidate genes for carotid plaque in Dominicans." Atherosclerosis **223**(1): 177-183.

Dumitrescu, D., et al. (2011). "Fully reversible pulmonary arterial hypertension associated with dasatinib treatment for chronic myeloid leukaemia." European Respiratory Journal **38**(1): 218-220.

Endelman, J. B. (2011). "Ridge regression and other kernels for genomic selection with R package rrBLUP." The Plant Genome **4**(3): 250-255.

Eu-Ahsunthornwattana, J., et al. (2014). "Comparison of methods to account for relatedness in genome-wide association studies with family-based data." PLoS Genet **10**(7): e1004445.

Fan, J. and R. Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American Statistical Association **96**(456): 1348-1360.

Friedman, J., et al. (2007). "Pathwise coordinate optimization." The Annals of Applied Statistics **1**(2): 302-332.

Friedman, J., et al. (2001). The elements of statistical learning, Springer series in statistics Springer, Berlin.

Friedman, J., et al. (2009). "glmnet: Lasso and elastic-net regularized generalized linear models." R package version **1**.

Friedman, J., et al. (2010). "A note on the group lasso and a sparse group lasso." arXiv preprint arXiv:1001.0736.

Friedman, J., et al. (2010). "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software **33**(1): 1.

Goldstein, B. A., et al. (2010). "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings." BMC genetics **11**(1): 49.

Groll, A. and G. Tutz (2014). "Variable selection for generalized linear mixed models by L 1-penalized estimation." Statistics and Computing **24**(2): 137-154.

Guymer, R., et al. (1993). "Benign intracranial hypertension in chronic myeloid leukemia." Australian and New Zealand journal of ophthalmology **21**(3): 181.

Hastie, T., et al. (2015). Statistical learning with sparsity: the lasso and generalizations, CRC Press.

Hong, H., et al. (2013). "Prehypertension is associated with increased carotid atherosclerotic plaque in the community population of Southern China." BMC cardiovascular disorders **13**(1): 20.

Houben, M., et al. (2004). "Hypertension as a risk factor for glioma? Evidence from a population-based study of comorbidity in glioma patients." Annals of oncology **15**(8): 1256-1260.

Howie, B. N., et al. (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet **5**(6): e1000529.

Ikram, M. K., et al. (2010). "Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo." PLoS Genet **6**(10): e1001184.

Jang, W. and J. Lim (2006). "PQL estimation biases in generalized linear mixed models." Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA: 05-21.

Javanmard, A. and A. Montanari (2013). "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression." arXiv preprint arXiv:1306.3171.

Javanmard, A. and A. Montanari (2014). "Confidence intervals and hypothesis testing for high-dimensional regression." The Journal of Machine Learning Research **15**(1): 2869-2909.

Kang, H. M., et al. (2010). "Variance component model to account for sample structure in genome-wide association studies." Nature genetics **42**(4): 348-354.

Kim, H. K., et al. (2013). "Osteogenic activity of collagen peptide via ERK/MAPK pathway mediated boosting of collagen synthesis and its therapeutic efficacy in osteoporotic bone by back-scattered electron imaging and microarchitecture analysis." Molecules **18**(12): 15474-15489.

Klein, R. J., et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-389.

Lange, K., et al. (2013). "Mendel: the Swiss army knife of genetic analysis programs." Bioinformatics **29**(12): 1568-1570.

Larson, N. B. and D. J. Schaid (2014). "Regularized Rare Variant Enrichment Analysis for Case ‐ Control Exome Sequencing Data." Genetic epidemiology **38**(2): 104-113.

Lazennec, G. and A. Richmond (2010). "Chemokines and chemokine receptors: new insights into cancer-related inflammation." Trends in molecular medicine **16**(3): 133-144.

Lea, A. J., et al. (2015). "A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data." PLoS Genet **11**(11): e1005650.

Lee, H. W., et al. (2008). "Berberine promotes osteoblast differentiation by Runx2 activation with p38 MAPK." Journal of Bone and Mineral Research **23**(8): 1227-1237.

Lee, S., et al. (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-775.

Lindstrom, M. J. and D. M. Bates (1988). "Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data." Journal of the American Statistical Association **83**(404): 1014-1022.

Lippert, C., et al. (2011). "FaST linear mixed models for genome-wide association studies." Nature methods **8**(10): 833-835.

Liu, L., et al. (2015). "Gene expression changes in human mesenchymal stem cells from patients with osteoporosis." Molecular medicine reports **12**(1): 981-987.

Mammen, E. and S. van de Geer (1997). "Penalized quasi-likelihood estimation in partial linear models." The Annals of Statistics: 1014-1035.

Mathieson, I. and G. McVean (2012). "Differential confounding of rare and common variants in spatially structured populations." Nature genetics **44**(3): 243-246.

McCulloch, C. E. (1997). "Maximum likelihood algorithms for generalized linear mixed models." Journal of the American Statistical Association **92**(437): 162-170.

McCulloch, C. E. and J. M. Neuhaus (2001). Generalized linear mixed models, Wiley Online Library.

Meier, L., et al. (2008). "The group lasso for logistic regression." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(1): 53-71.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(4): 417-473.

Mendelsohn, A. R. and J. Larrick (2013). "Dietary modification of the microbiome affects risk for cardiovascular disease." Rejuvenation research(ja).

Meng, L., et al. (2012). "Depression increases the risk of hypertension incidence: a meta-analysis of prospective cohort studies." Journal of hypertension **30**(5): 842-851.

Natarajan, B. K. (1995). "Sparse approximate solutions to linear systems." SIAM journal on computing **24**(2): 227-234.

Pearson, T. A. and T. A. Manolio (2008). "How to interpret a genome-wide association study." Jama **299**(11): 1335-1344.

Price, A. L., et al. (2010). "New approaches to population stratification in genome-wide association studies." Nature Reviews Genetics **11**(7): 459-463.

Qin, H., et al. (2010). "Interrogating local population structure for fine mapping in genome-wide association studies." Bioinformatics **26**(23): 2961-2968.

Rakitsch, B., et al. (2013). "A Lasso multi-marker mixed model for association mapping with population structure correction." Bioinformatics **29**(2): 206-214.

Rao, B. D. and K. Kreutz-Delgado (1999). "An affine scaling methodology for best basis selection." Signal Processing, IEEE Transactions on **47**(1): 187-200.

Robinson, G. K. (1991). "That BLUP is a good thing: the estimation of random effects." Statistical science: 15-32.

Schelldorfer, J., et al. (2014). "Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization." Journal of Computational and Graphical Statistics **23**(2): 460-477.

Simon, N., et al. (2013). "A sparse-group lasso." Journal of Computational and Graphical Statistics **22**(2): 231-245.

Sun, T. and C.-H. Zhang (2012). "Scaled sparse linear regression." Biometrika **99**(4): 879-898.

Thompson, E. and R. Shaw (1990). "Pedigree analysis for quantitative traits: variance components without matrix inversion." Biometrics: 399-413.

Thornton, T., et al. (2012). "Estimating kinship in admixed populations." The American Journal of Human Genetics.

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

Tripodi, G., et al. (1996). "Hypertension-associated point mutations in the adducin alpha and beta subunits affect actin cytoskeleton and ion transport." Journal of Clinical Investigation **97**(12): 2815.

van de Geer, S., et al. (2013). "On asymptotically optimal confidence regions and tests for high-dimensional models." arXiv preprint arXiv:1303.0518.

Vincent, M. and N. R. Hansen (2014). "Sparse group lasso and high dimensional multinomial classification." Computational Statistics & Data Analysis **71**: 771-786.

Voight, B. F. and J. K. Pritchard (2005). "Confounding from cryptic relatedness in case-control association studies." PLoS Genet **1**(3): e32.

Waldmann, P., et al. (2013). "Evaluation of the lasso and the elastic net in genome-wide association studies." Frontiers in genetics **4**(270).

Warriner, A. H. and K. G. Saag (2013). "Glucocorticoid-related bone changes from endogenous or exogenous glucocorticoids." Current Opinion in Endocrinology, Diabetes and Obesity **20**(6): 510-516.

Wu, M. C., et al. (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." The American Journal of Human Genetics **89**(1): 82-93.

Wu, T. T., et al. (2009). "Genome-wide association analysis by lasso penalized logistic regression." Bioinformatics **25**(6): 714-721.

XU, Z.-B., et al. (2012). "Representative of L1/2 Regularization among Lq Lq Lq (0< q⩽ 1) Regularizations: an Experimental Study Based on Phase Diagram." Acta Automatica Sinica **38**(7): 1225-1228.

Xu, Z., et al. (2012). "L1/2regularization: A thresholding representation theory and a fast solver." IEEE Transactions on neural networks and learning systems **23**(7): 1013-1027.

Xu, Z., et al. (2012). "regularization: A thresholding representation theory and a fast solver." Neural Networks and Learning Systems, IEEE Transactions on **23**(7): 1013-1027.

Yang, J., et al. (2014). "Advantages and pitfalls in the application of mixed-model association methods." Nature genetics **46**(2): 100-106.

Yi, N., et al. (2011). "Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects." PLoS genetics **7**(12): e1002382.

Yuan, M. and Y. Lin (2006). "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1): 49-67.

Zhang, C. and X. Chen (2009). "Smoothing projected gradient method and its application to stochastic linear complementarity problems." SIAM Journal on Optimization **20**(2): 627-649.

Zhang, L., et al. (2014). "FISH: fast and accurate diploid genotype imputation via segmental hidden Markov model." Bioinformatics: btu143.

Zhou, H., et al. (2010). "Association screening of common and rare genetic variants by penalized regression." Bioinformatics **26**(19): 2375-2382.

Zhou, Q., et al. (2014). "A reduction of the elastic net to support vector machines with an application to gpu computing." arXiv preprint arXiv:1409.1976.

Zhou, X. and M. Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies." Nature genetics **44**(7): 821-824.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2): 301-320.

BIOGRAPHY

Shaolong Cao was born in Xi’an, Shaanxi, China. I received my bachelor degree in Applied Mathematics in 2011 from Xi’an Jiaotong University in Xi’an, China. During my undergraduate study, I learned fundamental mathematics and computational science skills. After my graduation in Xi’an Jiaotong University, I decided to pursue my further academic study in Tulane University on Biomedical Engineering. Beginning at 2011 fall, I am working as a research assistant in Biomedical Engineering and the center for Bioinformatics and Genomics. My major research interest focuses on developing new sparse regularized regression based method for high-dimensional genetic data analysis. I am also investigating multi-omics data integration and high-dimensional Gaussian graphic model.