Winter 2014

Identification of genomic factors using family-based association studies

Libo Wang
Purdue University

Graduate School Form 30
(Updated 11/20/2014)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Libo Wang

Entitled
IDENTIFICATION OF GENOMIC FACTORS USING FAMILY-BASED ASSOCIATION STUDIES

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Dabao Zhang	Co-chair		
Min Zhang	Co-chair		
Mary E. Bock			
James C. Fleet			

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Dabao Zhang Co-chair

Min Zhang Co-chair

Approved by: Jun Xie 12/11/2014

Head of the Department Graduate Program Date

IDENTIFICATION OF GENOMIC FACTORS USING FAMILY-BASED ASSOCIATION STUDIES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Libo Wang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana

To my dear parents, Yanbin Wang and Li Pan,
and my beloved husband Yunfei Gao, my wonderful children Emma Gao and Andrew Gao.

ACKNOWLEDGMENTS

First and foremost, I thank Professor Dabao Zhang and Professor Min Zhang for their guidance and instruction during the time of my Ph.D. study. Their help to me has covered every aspect of my research life and they gave students the opportunity to get involved in the state-of-the-art research. Professor Dabao Zhang can always get my eyes open by pointing to new directions whenever I met problems. Professor Min Zhang's enthusiasm in the research has always made me excited about what I am doing. They are the people I went to talked to during my confusing and difficult time, and in some sense they are like mentors to me rather than research advisors. I feel fortunate to complete my PhD in their group.

I thank Prof. James Fleet for his advice and serving in my committee, it is Prof. Wanqing Liu's quick and clean thinking on many pharmacology related issues that really inspires me.

It has been my pleasure to be part of the Statistics Department at Purdue University. I am grateful to all the members in the department including professors, staffs and students. They have made a good working environment during my Ph.D. study. My special thank goes to Professor Rebecca Doerge and Professor Mary Ellen Bock for running such a fantastic department, Douglas G. Crabill and My Troung for their IT support, and Marian Duncan, Mary Roe, and Becca Miller for various help. I would also like to thank my colleagues, Yanzhu Lin, Vitara Pungpapong, Chen Chen, Qi Wang, Min Ren, and Chen Shi, all of them have always support me in many ways. There are several friends in the department I would like to mention the names here: Longjie Cheng, Simeng Qu, Zhaonan Sun, QiMing Huang, Pan Chao, Han Wu, Cheng Liu, Jeff Li, and Jingyi Zhu. They have made my life at Purdue an enjoyable experience.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

EMMA	efficient mixed-model association
EMMAX	efficient mixed-model association eXpedited
FBAT	family-based association test
FDR	false discovery rate
FP	false positives
fPOCRE	family-based penalized orthogonal component regression
GC	genomic control
GEE	generalized estimating equation
GEE-POCRE	penalized orthogonal component regression in the generalized estimating equation
GLM	generalized linear model
GWAS	genome-wide association study
MAF	minor allele frequency
ML	maximum likelihood
MLMM	multi-locus mixed model
PCA	principle component analysis
POCRE	penalized orthogonal component regression
REML	restricted maximum likelihood
ROADTRIPS	robust association-detection test for related individuals with population substructure
SA	structured association
SNP	single nucleotide polymorphism
TDT	transmission disequilibrium test
TR	true positives

ABSTRACT

Wang, Libo Ph.D., Purdue University, December 2014. Identification of Genomic Factors Using Family-Based Association Studies. Major Professor: Dabao Zhang and Min Zhang.

Genome-wide association studies become increasingly popular and important for detecting genetic associations of complex traits. However, it is well known that spurious associations could arise from statistical analysis without proper consideration of genetic relatedness of samples. Many methods have been proposed to guard against these spurious associations. Here we focus on multi-locus association studies of quantitative traits and the case-control status, and propose algorithms that take into consideration of genetic related samples to address possible confounding issues. As supervised dimension reduction methods, these algorithms performs well to conduct association studies with a large number of biomarkers but a relative small number of samples.

Recently, Linear mixed models have demonstrated its efficiency in GWAS of quantitative traits with multiple levels of sample structures. Most of the current mixed model based methods such as EMMA, EMMAX, and GEMMA, can be viewed as single-locus methods by testing each SNP separately. Complex traits, however, are known to be controlled by multiple loci, thus including multiple loci in the statistical model seems more appropriate. In the first part of my dissertation, we propose an algorithm that extends penalized orthogonal component regression to family-based association studies (fPOCRE) of continuous traits. While multiple loci can be investigated at the same time, the sample relatedness is modeled through the kinship matrix and the shared confounding effects are included as random effects in the linear mixed model. Our proposed algorithm simultaneously selects biomarkers and constructs their linear combinations as components which optimally account for variation

in traits. We compare fPOCRE with EMMAX, which is one of the most frequently used single-locus approach, and also compare it with MLMM, a recently developed multi-locus approach. Our simulation study demonstrates fPOCRE has promising performance over both EMMAX and MLMM in terms of higher power and fewer false positives when causal effects are from clusters of correlated SNPs. Real data are analyzed to illustrate the proposed approach and provide further comparisons.

Case-control association study is a widely used study design in genetic epidemiology and pharmacology and this study design is also susceptible to the potential confounding by sample structure. In the second part of my dissertation, we employ a multi-locus generalized estimation equation (GEE) model to study genetic associations of binary traits, capturing multiple levels of the sample structure with working correlation matrix. The kinship matrix is used to model the working correlation matrix, and the penalized orthogonal-components regression method is developed to build such a multi-locus GEE model (aka GEE-POCRE). GEE-POCRE is compared with gPOCRE, a multi-locus method that does not consider pedigree information, also compared with TDT, FBAT, and ROADTRIPS that are single-locus methods considering sample structure. In our simulation studies, GEE-POCRE demonstrates good performance in terms of protecting against spurious associations caused by the sample structure as well as having increased power.

1. INTRODUCTION

This chapter is served as an introduction by reviewing the problem of confounding in genetic association studies, which arises principally because of the population structure, family structure and cryptic relatedness. Moreover, it will also cover many of the existing solutions to this problem, such as genomic control, structured association, principle component analysis etc.

1.1 Spurious Associations Due to Genetic Relatedness of Samples

Genetic association studies are designed to identify genetic loci that contribute to the phenotypic outcomes of interest. The associations of interest are causal, finding loci whose different alleles have different effects. It is often that causal genetic loci are not directly genotyped in the study sample, in such cases, the associations can be found through closely link loci. Spurious associations are findings that are neither causal or nearby loci and they may arise when confounding factors are ignored. While population structure, family structure and cryptic relatedness describe different aspects of genetic relatedness among study subjects, they usually bring up confounding effects which, when ignored in association study, may result in misleading conclusions (Astle and Balding 2009).

A confounder is defined as a factor that is associated with both the exposure and the phenotype of interest. It is known that allele frequencies would vary among populations of different genetic ancestry and similarly, the trait of interest often varies among populations of different genetic ancestry. Therefore, SNPs that represent the sample genetic ancestry would become confounders that bias the association between the causal genetic factors and the phenotype of interest. For example, in a GWA study of the phenotype "eating more rice" (the phenotype of interest), the goal is

to find genetic markers (the exposure) that cause "eating more rice". The study samples are drawn from both north and south part of China and it is well known that southern Chinese eat more rice historically. Many alleles that are associated with southern Chinese (confounder) tend to show associations with the study phenotype, however, they are not the genetic factors that cause "eating more rice".

GWAS have been intensely developed recently, however, they still show limited successes. One of the many reasons is that spurious associations may occur due to structured samples. Even though McCarthy *et al.* (2008) concluded that the population structure should not have a big impact on the results of association studies when the cases and controls are well-matched, however, when the sample size increases so as to detect weak signals, even in populations with modest levels of structure, increasing false positives would be expected. The vulnerability of association studies to confounding effects caused by population structure has long been recognized. In a famous example, Knowler *et al.* (1988) found a significant association between an immunoglobulin haplotype and type II diabetes using samples drawn from native North Americans with some European ancestry, later the association disappeared when population structure was controlled.

Family structure refers to the genetic relatedness due to family structure among study samples and cryptic relatedness refers to the presence of close relatives in a sample of unrelated individuals (Price *et al.* 2010). While population structure describes a more distant common ancestry of large groups of individuals, cryptic relatedness refers to recent common ancestry among smaller groups of individuals (Astle and Balding 2009). Cryptic relatedness could cause spurious association in a way similar as population structure because of unmatched studies samples. Devlin and Roeder (1999) stated that cryptic relatedness would generate a more severe confounding problem than population structure if not properly handled.

1.2 Family-Based Association Test

Several methods have been proposed to conduct family-based association test so as to obviate the concerns about the confounding effects induced by structured samples. TDT compares the proportion of alleles transmitted versus the proportion of alleles not transmitted from the parents to the affected offspring (Spielman *et al.* 1993). Rabinowitz and Laird (2000) and Laird *et al.* (2000) proposed FBAT on the basis of the TDT method. It is a unified approach to family-based association tests, and can accommodate data with different combinations of family structure including nuclear families. FBAT applies to different phenotypes including case-control status, and can employ different genetic models including the additive model (Wu *et al.* 2005). Because of the independence among unlinked loci according to Mendel's second law, FBAT is immune to confounding due to sample structure.

1.3 Association Testing Methods for Population-Based Samples

Despite the ability of family-based linkage and association tests to handle the confounding issue, however, the power is limited by the sample size obtainable to detect relatively weak signals. Complex traits are known to be controlled by multiple loci with weak effects, many researchers, therefore, have turned to population-based association methods for its improved power to identify causal variants lying underneath the trait. Unfortunately, population-based study designs are susceptible to spurious associations due to hidden sample structure and many methods have been developed to tackle this problem. Here we discuss some commonly used association methods to identify causal genetic variants to the phenotype of interest while simultaneously controlling for the sample structure in population-based study designs.

1.3.1 Genomic Control

Genomic control (GC) proposed by Devlin and Roeder (1999) is one of the earliest methods that uses genomic information to correct for population structure. GC uses a set of random non-candidate markers to estimate an inflation factor λ and there is no population structure if λ equals one. GC scales the original χ^2 test statistic by the inflation factor λ, and the resultant test statistic follows a non-central χ^2 distribution. Though it is sufficient to adjust the test statistic by the estimated λ when the non-centrality parameter is small (Tiwari, 2008), this method suffers from loss of power when the non-centrality parameter is truly large. It is also noticeable that GC is a uniform adjustment to all the testing loci thus does not change their rankings. This makes GC less competitive compared with other association methods that explicitly account for population structure in the model itself. However, GC is a relatively easy method to implement and interpret and it requires a small number of markers.

1.3.2 Structured Association

Pritchard *et al.* (2000) introduced structured association (SA) that uses a set of random markers to estimate population structure of studying samples collected from unknown ancestry. This approach is also called an "island model" by assuming a fixed number of sub-populations/islands, and each individual is assigned to a cluster with a probability of a membership. More generally, assuming population admixture, SA can be viewed as a regression method by incorporating sub-populations as covariates. Similar to GC, SA can be effective using only hundreds of SNPs, however, unlike GC's simplicity, it can be slow to implement. Moreover, it is not appropriate to assume that only a limited number of ancestry groups exist for human populations.

1.3.3 Principle Component Analysis

Principle component analysis (PCA) is another popular tool to detect and adjust for underlying population admixture using genome-wide data (Price *et al.* 2006). PCA calculates continuous axes of genetic variation that maximize the variability between individuals, and reduces the data to a smaller number of dimensions. Similar to SA, PCs are included as covariates in a regression model to adjust for underlying population structure. Using top PCs to infer sub-population admixture is computational efficient. However, like SA, PCA only partially captures the multiple levels of the sample structure by assuming a limited number of ancestral populations.

1.3.4 Mixed Model Based Approaches

Single-Locus Mixed-Model

Yu *et al.* (2005) proposed to explicitly use kinship matrix in linear mixed models so as to model confounding effects induced by different levels of sample structure, including population structure, family structure and cryptic relatedness. A linear mixed model is composed of fixed effects and random effects: the effects of the candidate SNP, optional covariates, i.e. age and gender, are considered as fixed; confounding effects induced by sample structure are modeled as random effects and their correlations are described by kinship matrix (Prince *et al.* 2010). Linear mixed models are theoretically attractive but computationally intensive. The computation time increases with the cube of the number of individuals (Zhang *et al.* 2010). With this observation in mind, Kang *et al.* (2008) developed the efficient mixed model association (EMMA) by taking use of eigen-decomposition of the kinship matrix so as to facilitate global optimization of the likelihood function. Furthermore, Kang *et al.* (2010) proposed EMMA eXpedited (EMMAX) which is an approximate method that significantly reduces the computational time for analyzing large GWAS data sets. It

does not re-estimate variance parameters for every testing locus by assuming that effects of every SNP is very small.

Multi-Locus Mixed-Model

Complex traits are known to be controlled by several genetic factors. A gained power would be expected by testing all the loci simultaneously when compared with single-locus methods. With this observation in mind, Vincent *et al.* (2013) proposed a multi-locus stepwise mixed model regression (MLMM). The proposed MLMM algorithm does forward inclusion and stops when the genetic variance is zero, then performs backward elimination from the last model. Both extended Bayesian information criterion (eBIC) and multiple-Bonferroni criterion are suggested to choose the optimal model with eBIC slightly more stringent. Compared to single-locus mixed model methods, MLMM has more power and makes fewer false discoveries.

1.4 Variable Selection with High-Dimensional Data

1.4.1 Penalized Likelihood Methods

Nowadays with high-throughput technology, high-dimensional data are becoming increasingly common in many areas of disease related studies. The goal of high-dimensional data analyses is to uncover the underlying structure that regulates the trait of interest. One of the many obstacles associated with analyzing large p small n data is how to select important variables that have non-zero effects. Many variable selection techniques have been developed to tackle this problem.

One way of doing variable selection in high-dimensional data is using the penalized likelihood method which does variable selection and estimation simultaneously. Letting $\ell(\beta)$ be the negative log-likelihood function, the maximum likelihood estimator is obtained as

$$\hat{\beta} = \arg\min_{\beta} \ell(\beta). \tag{1.1}$$

A penalty function $P_\lambda(\beta)$ is added to the objective function to obtain sparse estimators and we have the following new objective function

$$\hat{\beta} = \arg\min_{\beta} \left(\ell(\beta) + P_\lambda(\beta) \right), \tag{1.2}$$

where $\lambda \geq 0$ is the tuning parameter, with larger λ leading to a more sparse model. LASSO proposed by Tibshirani (1996) is probably the most well known penalty function which puts an ℓ_1-norm on the coefficients and can be denoted as $P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$. Even though LASSO is relative easy to compute, however, it is known for lacking of grouped ability, only able to select one variable from a group of highly correlated predictors (Zhang *et al.* 2008). It is also known that LASSO estimates are biased and not consistent. Over the years, many methods have been proposed that use various penalty function to improve the performance, such as adaptive LASSO, smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP) etc.

Zou (2006) proposed adaptive LASSO which puts weights $\hat{w}$ on the coefficients. The penalty function is expressed as $P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|$, where $\hat{w}_j = 1/|\hat{\beta}_j^c|^\gamma$ and λ is the tuning parameter. $\hat{\beta}^c$ is any consistent estimator. Estimates from ordinary least square can be used if the sample size is larger than the number of variables, otherwise, ridge regression estimates are suggested.

Developed by Fan and Li (2001), SCAD penalty is defined as

$$P_\lambda(\beta) = \sum_{j=1}^{p} P_{\lambda,j}(\beta_j; a),$$

where, with $a > 2$ and $\lambda > 0$,

$$P_{\lambda,j}(\beta_j; a) = \begin{cases} \lambda|\beta_j|, & |\beta_j| < \lambda; \\ -(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)/[2(a-1)], & \lambda < |\beta_j| \leq a\lambda; \\ (a+1)\lambda^2/2, & |\beta_j| > a\lambda. \end{cases}$$

SCAD has the ability to produce unbiased and sparse estimators, moreover, it enjoys oracle properties.

1.4.2 Supervised Dimension Reduction Methods

An alternative way in analyzing large p small n problem is to reduce the dimension by constructing components. The key is using a low-dimensional space to represent the information contained in the original high-dimensional space. Partial least square (PLS) method (Wold *et al.* 1984) is an efficient supervised dimension reduction method, however, it does not produce sparsely estimated coefficients. Zhang et al. (2008) developed Penalized Orthogonal Components Regression (POCRE) to enable PLS for variable selection.

As shown in Zhang *et al.* (2008), POCRE seeks to construct a sequence of penalized orthogonal components. The loading of each constructed component is obtained by maximizing the correlation between the response and the component, and a penalty function is imposed to help identify sparse predictors for each component. POCRE is easy to implement and fast to compute. Moreover, unlike LASSO, POCRE has the ability to group highly correlated predictors. The model is expressed as

$$Y \quad = \quad \mathbf{X}\beta + \epsilon, \tag{1.3}$$

where Y is the response vector, $\mathbf{X}$ is the $n \times p$ design matrix, β is the $p \times 1$ fixed effect vector, and ϵ is the residual vector with $Var(\epsilon) = \sigma_e^2$. Moreover, ϵ is assumed to be independent of the columns of the design matrix $\mathbf{X}$. POCRE starts with $\mathbf{X}_1 = \mathbf{X}$ and proceeds to build a sequence of orthogonal components. Assuming the first $k-1$ orthogonal components have been constructed, we proceed to build the kth component $\mathbf{X}_k\omega_k$. The loading ω_k is calculated as $\omega_k = \frac{\nu}{\|\nu\|}$ where ν minimizes $-2\nu^T\mathbf{X}_k^T YY^T\mathbf{X}_k\alpha + \|\nu\|^2 + P_\lambda(\nu)$, subject to $\|\alpha\| = 1$, $P_\lambda(\nu)$ is a penalty function with tuning parameter λ and the current implementation includes empirical bayes thresholding (Johnstone and Silverman 2004), L_1 Penalty (Tibshirani 1996), SCAD (Fan and Li 2001), and MCP (Zhang 2010).

1.5 Motivation for Dissertation Research

The Multi-Locus Mixed Model (MLMM) has drawn a lot of attention since it first appeared. Its very nature of being a multi-locus testing method separates itself from many other existing single-locus methods, i.e. EMMA, EMMAX, GEMMA. It is noted by the authors that LASSO-type penalization methods are worth to investigate since stepwise regression can not explore all the different combinations of the variable spaces (Vincent *et al.* 2012). However, LASSO-type penalization methods is known for lacking grouping property, and this hinders the application of LASSO to many biological studies due to sharing pathways of many genetic factors (Zhang *et al.* 2008). An approach that uses a variable selection algorithm suitable for structured predictors in conjunction with linear mixed model could potentially improves the performance.

The linear mixed model that explicitly uses the kinship matrix may help solve the confounding problem of the association study with quantitative traits. However if it concerns case-control family data, which is a common study design in genome-wide association studies, how to simultaneously avoid spurious associations and improve power becomes a challenge. Current methods that based on TDT-type of association tests are essentially single-locus analyses, making them less powerful. With the similar insights of authors of MLMM, an advanced statistical method that models multiple loci simultaneously to fully utilize the information contained in the data would potentially make great improvements over existing methods.

The rest of the dissertation is organized as following. Chapter 2 describes our proposed algorithm fPOCRE which is a penalized multi-locus method that constructs orthogonal components assuming a linear mixed model. The proposed algorithm is compared with other popular methods in both simulation studies and real data analyses. It is followed by Chapter 3, which describes our proposed GEE-POCRE that does the variable selection through a penalization function under the generalized estimating equation (GEE) model. GEE has long been recognized in longitudinal study of case-control analyses to handle correlated data. The proposed algorithm

has great potential to prevent spurious associations due to sample structure in the case-control study design. Finally, Chapter 4 summarizes my research and lists some future potential researches.

2. A MULTI-LOCUS METHOD FOR FAMILY-BASED GENOME-WIDE ASSOCIATION STUDY

In this chapter, we describe the details of the proposed family-based penalized orthogonal components regression (fPOCRE) in genome-wide association studies with quantitative traits. While multiple loci can be investigated at the same time, the sample relatedness is modeled through the kinship matrix and the shared confounding effects are included as random effects in the linear mixed model. Our proposed algorithm simultaneously selects biomarkers and constructs their linear combinations as components which optimally account for variation in traits. We compare fPOCRE with two other methods based on linear mixed models, i.e., EMMAX, a single-locus approach, and MLMM, a recently developed multi-locus approach. Our simulation study demonstrates that fPOCRE has promising performance over these two popular methods in terms of improved power and low false positives when the causal effects are from clusters of correlated SNPs. Real data analyses are used to illustrate the proposed approach and provide further comparisons.

We start this chapter with the motivation of our novel algorithm. Section 2.2 describes our proposed fPOCRE for the linear mixed model. The full details on the algorithm are provided in Section 2.3. Section 2.4 and Section 2.5 contain the results of simulations and real data analyses respectively.

2.1 Introduction and Motivation

With the recent advances in high-throughput biotechnologies, genome-wide association studies (GWAS) are becoming more and more popular in analyzing underlying genetics of both disease status and quantitative traits. However with the presence of population stratification, and additional complexities such as family structure or

cryptic relatedness, GWAS could produce false signals if not handled properly (Yu *et al.* 2005). Several methods have been proposed to address this problem, for example, Transmission Disequilibrium Test (TDT) (Spielman *et al.* 1993), and Family-Based Association Test (FBAT) (Rabinowitz and Laird, 2000; Laird *et al.* 2000). FBAT covers TDT as a special case and is known for its ability to prevent spurious association due to population structure. However, it only applies to certain study design which is not always attainable and usually presents a small sample compared to a large sample in a population study design.

A widely used approach to detect the existence of population structure is to compute the genomic control parameter λ_{GC} (Devlin and Roeder 1999). This method could also be used to correct for population structure. However, it usually does not maximize the power of detecting true associations, and also does not change the rank of the detected signals. Other approaches, including structured association (Pritchard *et al.* 2000) and principle component analysis (Price *et al.* 2006), are only able to correct for population structure, but not family structure and cryptic relatedness.

Recently linear mixed model has been demonstrated as a way to simultaneously address confounding due to population structure, family structure and cryptic relatedness (Yu et al. 2005). The random effects in the linear mixed model can be used to model the sample relationship, which is described by the kinship matrix. EMMA (Kang *et al.* 2008) and EMMAX (Kang *et al.* 2010) are the two frequently used algorithms that account for the confounding factors with random effects in mixed model, with EMMAX being more computationally efficient. Even though these methods have been shown to have improved ability to reduce both false positives and false negatives, they are essentially still single-locus methods. It loses power when comparing to multi-locus methods. With this observation in mind, Vincent *et al.*(2013) proposed a stepwise multi-locus approach (MLMM) based on a linear mixed model. However, Breiman (1996) showed that classical stepwise regression is unstable due to the reason that modifying a single observation could produce an entirely different

model, and collinearity between variables makes this problem even worse because the correlation among the predictors could affect the order of selected signals.

Utilizing sparsity principle, penalized regression methods, especially those based on the LASSO algorithm (Tibshirani 1996), have received a lot of attention in recent years. It gains its popularity due to the fact these methods could simultaneous do variable selection and coefficient estimation in the large p small n scenario. However it is also known that LASSO is lack of the ability to select a group of correlated causal predictors (Zou and Hastie 2005). An alternative way to analyzing the large p small n problem is to reduce the predictor dimension by constructing components, such as partial least square (PLS) method (Wold *et al.* 1984) that is a supervised dimension reduction technique. Penalized orthogonal components regression (POCRE)(Zhang *et al.* 2009) is one of those notable PLS-based penalization methods which can effectively handle highly correlated covarites in high dimensional analyses. With these observations in mind, we propose an algorithm that extends penalized orthogonal components regression in the context of linear mixed model for family-based association studies (fPOCRE). This hybrid algorithm could take the advantage of both the linear mixed model and the penalized regression.

2.2 Extending the Penalized Orthogonal Components Regression to Family-Based Genome-Wide Association Studies

A linear mixed model can be expressed as (Pinheiro and Bates 2000, Ch.2)

$$Y \;=\; \mathbf{X}\beta + \mathbf{Z}u + \epsilon, \qquad \epsilon \sim N(0, \sigma_e^2 I_n), \tag{2.1}$$

where Y is an $n \times 1$ column vector, $\mathbf{X}$ is an $n \times p$ design matrix of the fixed effects, β is a $p \times 1$ vector representing coefficients of fixed effects, Z is the design matrix of the random effects, u is an $n \times 1$ random effect vector, and ϵ is an $n \times 1$ vector representing residual effect. The random effect u and residual effect ϵ are assumed to be independent. We further assume that

$$u \sim N\left(0, \tau\sigma_e^2\mathbf{K}\right),$$

where $\mathbf{K}$ is the $n \times n$ kinship matrix estimated from either pedigree or genomic information, τ is the ratio between the genetic variance σ_u^2 and the residual variance σ_e^2.

Note that $\mathbf{Z}$ is an $n \times n$ incidence matrix mapping each observed phenotype to one of the observations, therefore it is an identity matrix $\mathbf{I}$ in the model. Then the proposed linear mixed model can be re-expressed as

$$Y = \mathbf{X}\beta + u + \epsilon = \mathbf{X}\beta + \epsilon^*, \tag{2.2}$$

where $\epsilon^* = u + \epsilon$. As it is the sum of two independent multivariate normal vectors, it is independently distributed as multivariate normal with mean 0 and variance-covariance matrix $\sigma_e^2\mathbf{V}$, where $\mathbf{V} = \mathbf{I} + \tau\mathbf{K}$. Therefore we have Y following a multivariate normal distribution with mean $\mathbf{X}\beta$ and the variance-covariance matrix $\sigma_e^2\mathbf{V}$, and that is $Y \sim N(\mathbf{X}\beta, \sigma_e^2\mathbf{V})$.

The corresponding likelihood function could be written as

$$L(\beta, \tau, \sigma_e^2) = (2\pi\sigma_e^2)^{-n/2} \exp\left(\frac{(Y - \mathbf{X}\beta)^T\mathbf{V}^{-1}(Y - \mathbf{X}\beta)}{-2\sigma_e^2}\right)|\mathbf{V}|^{-1/2}. \tag{2.3}$$

For a given value of τ, the values of β and σ_e^2 that maximize the likelihood function can be written as

$$\hat{\beta}(\tau) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}Y, \tag{2.4}$$

$$\hat{\sigma_e^2}(\tau) = \frac{(Y - \mathbf{X}\hat{\beta})^T\mathbf{V}^{-1}(Y - \mathbf{X}\hat{\beta})}{n}. \tag{2.5}$$

Using these expressions, the profiled log-likelihood can be derived as

$$\log L(\tau) = -\frac{n}{2}\log(2\pi\hat{\sigma_e^2}) - \frac{n}{2} - \frac{1}{2}\log|\mathbf{V}|. \tag{2.6}$$

Hence, the MLE of τ is found by maximizing the objective function (2.6) w.r.t. τ, and the MLEs of β and σ_e^2 are estimated according to (2.4) and (2.5). This iterative procedure works well when the sample size n is reasonably larger than the number of predictors p, however, with nowadays technology, the large p small n issue is quite

common, which makes finding MLE of β an ill-conditioned problem. Therefore, it is nature for us to think of using a penalization step to help produce a legitimately estimated β.

With an initial estimation of τ, we define

$$\tilde{Y} = \mathbf{V}^{-1/2} Y, \tag{2.7}$$

and

$$\tilde{\mathbf{X}} = \mathbf{V}^{-1/2} \mathbf{X}. \tag{2.8}$$

Then components of $\tilde{Y}$ are independent, specifically $\tilde{Y} \sim N(\tilde{\mathbf{X}}\beta, \sigma_e^2 \mathbf{I}_n)$. POCRE can be henceforth applied to construct a sequence of orthogonal components for the purpose of estimating β. Starting with $\mathbf{X}_1 = \tilde{\mathbf{X}}$, $Y_1 = \tilde{Y}$, we build the first orthogonal component $\mathbf{X}_1\omega_1$, where ω_1 is the leading eigenvector of $cov(Y_1, \mathbf{X}_1)^T cov(Y_1, \mathbf{X}_1)$. Assuming that the first $k-1$ components are constructed, now we proceed to find the k-th component, first we remove $\mathbf{X}_{k-1}\omega_{k-1}$ from $\mathbf{X}_{k-1}$ and define $\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{X}_{k-1}\omega_{k-1}$ so that $\mathbf{X}_k$ is orthogonal to the previously constructed component. In a similar manner, Y_k is calculated by removing $\mathbf{X}_{k-1}\omega_{k-1}$ from Y_{k-1} so that Y_k is uncorrelated to $\mathbf{X}_{k-1}\omega_{k-1}$. Then ω_k is calculated as the leading eigenvector of $cov(Y_k, \mathbf{X}_k)^T cov(Y_k, \mathbf{X}_k)$. This procedure continues until there is no more correlation between the residual Y_k and $\mathbf{X}_k$. To enforce sparsity, the loadings are estimated as $\omega_k = \frac{\nu}{\|\nu\|}$ where ν minimizes $-2\nu^T \mathbf{X}_k^T Y Y^T \mathbf{X}_k \alpha + \|\nu\|^2 + P_\lambda(\nu)$, subject to $\|\alpha\| = 1$, here $P_\lambda(\nu)$ is a penalty function with tuning parameter λ. Currently, the penalty functions that have been implemented are L_1, SCAD, MCP, EBT and EBTZ which is EBT with a z-transformation. This whole procedure provides us a data-driven sparse estimate of β,

$$\hat{\beta} = \sum_{j=1}^{l} \zeta_j \omega_j Q_j, \tag{2.9}$$

where $\zeta_1 = I_{p\times p}$, $\zeta_{j+1} = \zeta_j(I - \omega_j P_j)$, $P_j = \eta_j^T \mathbf{X}_j / \eta_j^T \eta_j$, $Q_j = \eta_j^T Y_j / \eta_j^T \eta_j$, $\eta_j = \mathbf{X}_j \omega_j$, and $\omega_j = \frac{\nu}{\|\nu\|}$.

The residuals R are then obtained as $R = Y - \mathbf{X}\hat{\beta} = u + \epsilon = \epsilon^* \sim N(0, \sigma_e^2 \mathbf{V})$, which is a random intercept model. As we described previously, τ can be estimated by maximizing the profiled log-likelihood function in (2.6). $\mathbf{V}$ is then updated according to this newly estimated τ. We iteratively update β and τ, and stop whenever τ converges. The final estimate of β is calculated by constructing a sequence of penalized orthogonal components using the converged τ. Next section describes our proposed algorithm in details.

2.3 The Algorithm

Without loss of generality, we assume that both $\mathbf{X}$ and Y are centered, and L_1 penalty is applied to enforce sparsity. Algorithms using other penalty functions are similar. First of all, we eigen-decompose the kinship matrix and have $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where columns of $\mathbf{U}$ are the eigenvectors and $\mathbf{D}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{K}$. For each fixed tuning parameter λ in L_1 penalty, the fPOCRE algorithm proceeds as follows,

0. Set initial value $\tau^{(0)} = 0$ if this is the first λ being considered, otherwise, let $\tau^{(0)}$ be the one estimated from previous λ. Let k represents the number of iterations, and it starts with $k = 0$;

1. With $\tau = \tau^{(k)}$, obtain β by constructing penalized orthogonal components in regressing $\tilde{Y} = (I + \tau\mathbf{D})^{-1}\mathbf{U}^T Y$ against $\tilde{\mathbf{X}} = (\mathbf{I} + \tau\mathbf{D})^{-1}\mathbf{U}^T\mathbf{X}$;

2. Update $R = \tilde{Y} - \tilde{\mathbf{X}}\hat{\beta}$, and obtain $\tau^{(k+1)}$ by maximizing the following profiled log-likelihood,

$$\tau = \arg\min_{\tau} \left\{ \frac{n}{2} \log\left(\frac{R^T(\mathbf{I} + \tau\mathbf{D})^{-1}R}{n}\right) + \frac{1}{2}\log|\mathbf{I} + \tau\mathbf{D}| \right\}; \tag{2.10}$$

3. Iterate 1-2 until $\tau^{(k)}$ converges.

The proposed algorithm starts with $\lambda = 1$ with increment δ and continues until no features are being selected, then decreases from $\lambda = 1$ with decrement δ, the algorithm continues until the number of selected features are more than $n/\log n$.

Notice that the algorithm described above requires the choice of the regularization parameter λ, extended BIC (eBIC) proposed by Chen and Chen (2008) is suggested to help us choose the appropriate model for its effectiveness of variable selection in the large p small n problem. The eBIC algorithm is defined as

$$\text{eBIC}_\gamma = -2\log L + k\log n + 2\gamma\log\binom{p}{k}, \tag{2.11}$$

where L is the likelihood function of the linear mixed model, and k is the number of features being selected. We elicit γ from 0 to 1, with EBIC_0 exactly the BIC.

2.4 Simulation Studies

To maintain the unique structure of the genetic information associated with real pedigree data, the simulated data were generated by adding effects to the real *Arabidopsis* genetic information (Horton *et al.* 2012). Our proposed fPOCRE is compared with EMMAX which is an efficient single-locus mixed model approach that takes care of the relatedness among samples, as well as MLMM, a multi-locus mixed model approach that uses kinship matrix to capture the sample relatedness. We set up two simulation scenarios, and simulate 100 data sets in each case. In the first simulation scenario, a set of phenotypes is simulated by adding fixed effects to a group of 12 SNPs that are mildly to highly correlated. In the second simulation scenario, phenotypes are simulated from three different linkage groups, with 11 SNPs in the first group, seven SNPs in the second group, and 10 SNPs in the last group.

Case I. A Cluster of Mildly to Highly Correlated SNPs

We select a group of 12 SNPs, which are mildly to highly correlated. The minimum pairwise correlation among these 12 SNPs is 0.31 and the highest pair is 0.95. The phenotypic values are simulated by assuming fixed effects of these 12 SNPs, and random effects which are correlated according to the kinship matrix. Specifically, the underlying true model is assumed as

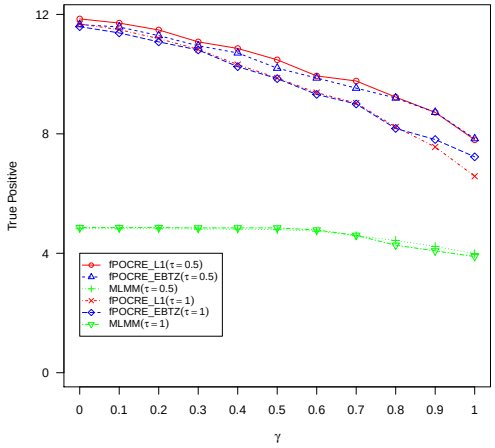
$$Y = \sum_{j=1}^{12} X_j + u + \epsilon, \qquad \epsilon \sim N(0,5), \quad u \sim N(0, 5\tau\mathbf{K}),$$

where $\mathbf{K}$ is the kinship matrix estimated from the genetic information, and τ is the ratio between the genetic variation and the residual variation. τ represents different levels of the signal-to-noise ratio, and is pre-specified at $\tau = 0.5$, 1, 5, 10, 20, 50, and 100 respectively.

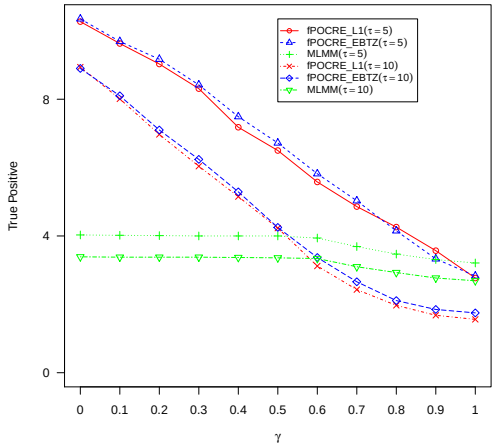
When it comes to fPOCRE, the result depends on the type of penalization functions used, e.g. L_1, SCAD, MCP, EBT, EBTZ penalty, and it also depends on the parameter γ used in eBIC to select the optimal model. The γ choices we consider for fPOCRE in our simulation studies are from 0 to 1 with step size at 0.1. Although MLMM only implements eBIC with $\gamma = 1$, we modified MLMM to allow eBIC taking different γ values so that we could have a thorough comparison. False Discovery Rate (FDR) is employed to adjust multiple testings in EMMAX, and two sets of EMMAX results are reported, one by controlling FDR at 0.05 and the other one at 0.1.

Since fPOCRE with L_1, SCAD and MCP penalty perform similarly, and fPOCRE with EBT and EBTZ penalty work similarly (results not shown here), we henceforth only show the results of fPOCRE with L_1 and EBTZ penalty.

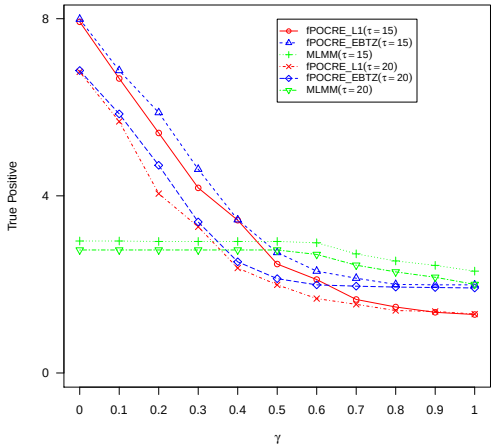
Figure 2.1 and Figure 2.2 compare results of fPOCRE(L_1), fPOCRE(EBTZ), and MLMM with different choices of γ. Figure 2.1 illustrates how the mean true positives of fPOCRE and MLMM changes with γ in eBIC increases, and Figure 2.2 illustrates how the mean false positives of fPOCRE and MLMM changes with γ in eBIC increases. Knowing that the larger the γ in eBIC, the more stringent model is, therefore less signals are found. We observe that fPOCRE with a smaller γ performs better than the one with a larger γ. Among all the γ used in the simulation study, fPOCRE with $\gamma = 0.1$ performs the best. On the other hand, MLMM with a larger γ outperforms the one with a smaller value, this is probably the reason that MLMM algorithm uses $\gamma = 1$ as the only value in their eBIC implementation of choosing the best model. Additionally, fPOCRE has a better performance than MLMM in terms of having a higher number of true detections and a lower number of false selections. When γ is smaller, the difference between fPOCRE and MLMM is bigger.



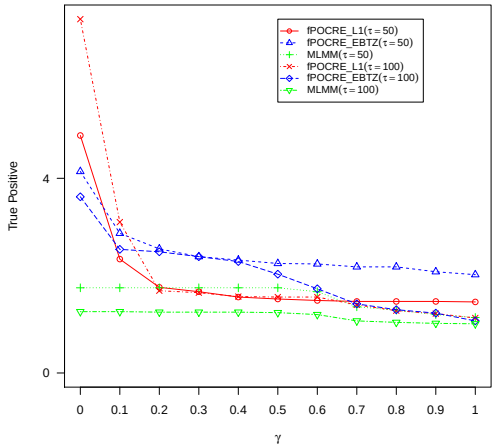
(a) True positives: $\tau = 0.5$ & $\tau = 1$



(b) True positives: $\tau = 5$ & $\tau = 10$

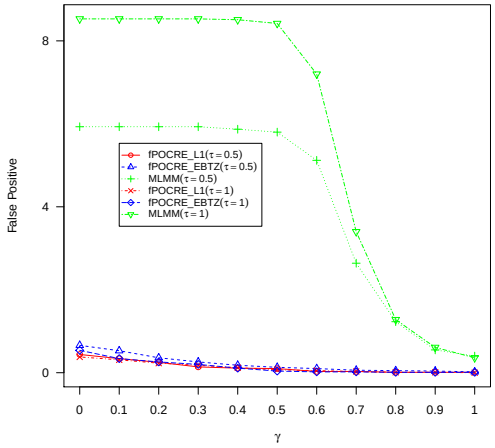


(c) True positives: $\tau = 15$ & $\tau = 20$

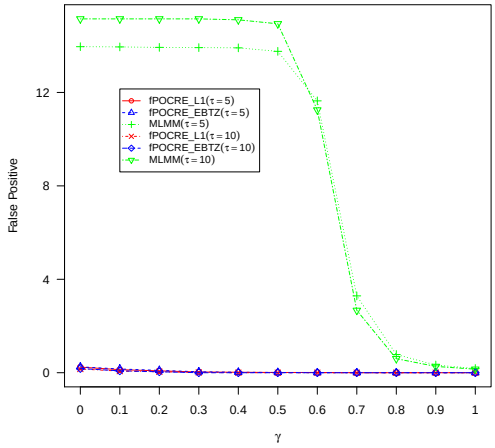


(d) True positives: $\tau = 50$ & $\tau = 100$

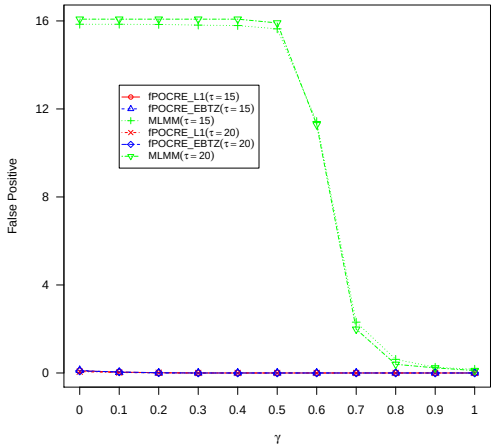
Figure 2.1.: True positives of fPOCRE(L_1), fPOCRE(EBTZ) and MLMM algorithms in the simulation study of a cluster of 12 SNPs



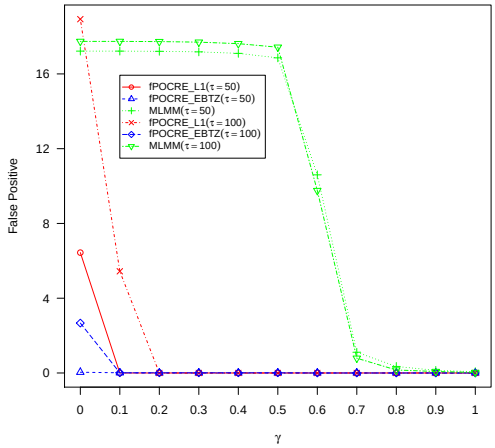
(a) False positives: $\tau = 0.5$ & $\tau = 1$



(b) False positives: $\tau = 5$ & $\tau = 10$



(c) False positives: $\tau = 15$ & $\tau = 20$



(d) False positives: $\tau = 50$ & $\tau = 100$

Figure 2.2.: False positives of fPOCRE(L_1), fPOCRE(EBTZ) and MLMM algorithms in the simulation study of a cluster of 12 SNPs

Table 2.1.: Performance comparison in analyzing a cluster of 12 mild to high correlated SNPs. Reported are the mean true positives (TP) across 100 simulated data sets with standard errors presented in the parentheses.

Method	$\tau = 0.5$	$\tau = 1$	$\tau = 5$	$\tau = 10$
fPOCRE$(L_1)_{\gamma=0.1}$	11.71(0.07)	11.49(0.09)	9.63(0.15)	8.01(0.18)
fPOCRE$(L_1)_{\gamma=1}$	7.80(0.23)	6.58(0.24)	2.77(0.20)	1.56(0.11)
fPOCRE(EBTZ)$_{\gamma=0.1}$	11.58(0.07)	11.38(0.11)	9.69(0.16)	8.10(0.18)
fPOCRE(EBTZ)$_{\gamma=1}$	7.84(0.22)	7.23(0.23)	2.83(0.20)	1.75(0.10)
MLMM$_{\gamma=0.1}$	4.82(0.11)	4.86(0.10)	4.02(0.07)	3.38(0.06)
MLMM$_{\gamma=1}$	3.99(0.07)	3.89(0.06)	3.21(0.05)	2.69(0.06)
EMMAX(FDR=0.05)	12.00(0.00)	12.00(0.00)	12.00(0.00)	12.00(0.00)
EMMAX(FDR=0.1)	12.00(0.00)	12.00(0.00)	12.00(0.00)	12.00(0.00)
Method	$\tau = 15$	$\tau = 20$	$\tau = 50$	$\tau = 100$
fPOCRE$(L_1)_{\gamma=0.1}$	6.65(0.23)	5.68(0.22)	2.34(0.15)	3.10(0.34)
fPOCRE$(L_1)_{\gamma=1}$	1.32(0.07)	1.33(0.06)	1.46(0.08)	1.13(0.07)
fPOCRE(EBTZ)$_{\gamma=0.1}$	6.83(0.20)	5.85(0.22)	2.87(0.20)	2.54(0.15)
fPOCRE(EBTZ)$_{\gamma=1}$	1.99(0.13)	1.92(0.13)	2.02(0.12)	1.07(0.10)
MLMM$_{\gamma=0.1}$	2.98(0.07)	2.78(0.07)	1.75(0.06)	1.26(0.05)
MLMM$_{\gamma=1}$	2.30(0.05)	2.01(0.05)	1.14(0.03)	1.01(0.01)
EMMAX(FDR=0.05)	12.00(0.00)	12.00(0.00)	11.98(0.14)	10.99(1.24)
EMMAX(FDR=0.1)	12.00(0.00)	12.00(0.00)	11.99(0.10)	11.28(1.06)

Table 2.2.: Performance comparison in a cluster of 12 mild to high correlated SNPs. Reported are the mean false positives (FP) across 100 simulated data sets with standard errors presented in the parentheses.

Method	$\tau = 0.5$	$\tau = 1$	$\tau = 5$	$\tau = 10$
fPOCRE(L_1)$_{\gamma=0.1}$	0.34(0.06)	0.31(0.06)	0.13(0.04)	0.07(0.03)
fPOCRE(L_1)$_{\gamma=1}$	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
fPOCRE(EBTZ)$_{\gamma=0.1}$	0.53(0.08)	0.34(0.06)	0.16(0.04)	0.08(0.03)
fPOCRE(EBTZ)$_{\gamma=1}$	0.02(0.01)	0.01(0.01)	0.00(0.00)	0.00(0.00)
MLMM$_{\gamma=0.1}$	5.93(0.46)	8.53(0.45)	13.95(0.24)	15.15(0.17)
MLMM$_{\gamma=1}$	0.40(0.06)	0.36(0.06)	0.18(0.04)	0.15(0.04)
EMMAX(FDR=0.05)	43.96(3.83)	43.52(3.85)	38.14(4.12)	32.78(4.43)
EMMAX(FDR=0.1)	53.55(5.97)	52.70(5.82)	46.04(5.26)	39.55(5.97)
Method	$\tau = 15$	$\tau = 20$	$\tau = 50$	$\tau = 100$
fPOCRE(L_1)$_{\gamma=0.1}$	0.04(0.02)	0.02(0.01)	0.00(0.00)	5.44(1.37)
fPOCRE(L_1)$_{\gamma=1}$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
fPOCRE(EBTZ)$_{\gamma=0.1}$	0.03(0.02)	0.01(0.01)	0.01(0.01)	0.01(0.01)
fPOCRE(EBTZ)$_{\gamma=1}$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
MLMM$_{\gamma=0.1}$	15.85(0.11)	16.08(0.10)	17.22(0.07)	17.74(0.05)
MLMM$_{\gamma=1}$	0.17(0.05)	0.11(0.03)	0.07(0.03)	0.03(0.02)
EMMAX(FDR=0.05)	28.78(4.64)	25.59(3.94)	13.98(4.57)	5.78(3.58)
EMMAX(FDR=0.1)	34.57(5.80)	30.90(5.63)	17.67(5.44)	8.29(4.95)

The true positives results of a cluster of 12 correlated SNPs simulation study are summarized in Table 2.1 and the false positives results are in Table 2.2. We present the results of fPOCRE using L_1 and EBTZ penalty with $\gamma = 0.1$ and $\gamma = 1$, MLMM results with $\gamma = 0.1$ and $\gamma = 1$, in additional to the results above, we also include two sets of EMMAX results at different FDR level to adjust for multiple comparisons. It shows that EMMAX produces really high false positives across all the τ considered compared with fPOCRE and MLMM. The number of false positives decreases when the ratio between the genetic variation and the residual variation increases. On the other hand, MLMM has low power compared with fPOCRE and EMMAX, and is only able to find at most one third of the total signals in this simulation study. In all of different τ settings. fPOCRE is the one with both a high detection rate and low false positives compared to its two competitors. Among fPOCRE algorithms itself, models with different penalization functions, fPOCRE$(L_1)_{\gamma=0.1}$ gives slightly higher TP than fPOCRE(ebtz)$_{\gamma=0.1}$ when the ratio between genetic variation and residual variation is relatively small, such as 0.5 and 1. With this ratio getting larger, fPOCRE(ebtz)$_{\gamma=0.1}$ performs better in terms of having higher TP. Furthermore, fPOCRE$(L_1)_{\gamma=0.1}$ generally has similar FP or a bit lower FP when compared with fPOCRE(ebtz)$_{\gamma=0.1}$ except the case when true $\tau = 100$.

Case II. Three Linkage Groups

We select a total of 28 SNPs in three different linkage groups: the first group has 11 SNPs, the second group has seven SNPs, and the last group has 10 SNPs. The phenotypic values are simulated by assuming fixed effects of these 28 SNPs, and random effects which are correlated according to the kinship matrix. Specifically, the underlying true model is assumed as

$$Y = \sum_{j=1}^{28} X_j + u + \epsilon, \qquad \epsilon \sim N(0, 5), \quad u \sim N(0, 5\tau\mathbf{K}),$$

where $\mathbf{K}$ is the kinship matrix estimated from genetic information, τ is the ratio between genetic variation and residual variation. While τ represents different levels

of the signal-to-noise ratio, it is pre-specified at $\tau = 0.5$, 1, 5, 10, 20, 50, and 100 respectively.

Table 2.3.: Performance comparison in the three linkage groups simulation study. Reported are the mean true positives (TP) across 100 simulated data sets with standard errors presented in the parentheses.

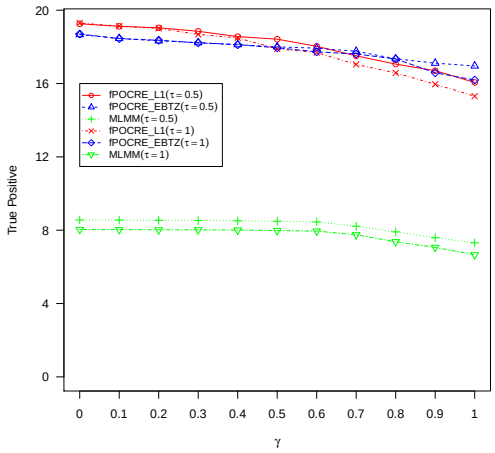
Method	$\tau = 0.5$	$\tau = 1$	$\tau = 5$	$\tau = 10$
fPOCRE(L1)$_{\gamma=0.1}$	19.12(0.75)	19.13(0.82)	17.81(1.32)	16.27(1.77)
fPOCRE(L1)$_{\gamma=1}$	16.06(2.13)	15.31(3.26)	4.63(5.61)	1.50(0.91)
fPOCRE(EBTZ)$_{\gamma=0.1}$	18.46(0.97)	18.45(0.90)	17.44(1.10)	16.26(1.58)
fPOCRE(EBTZ)$_{\gamma=1}$	16.96(1.63)	16.19(2.78)	4.81(5.61)	2.16(2.01)
MLMM$_{\gamma=0.1}$	8.55(2.39)	8.03(2.04)	6.47(1.46)	5.31(1.21)
MLMM$_{\gamma=1}$	7.31(1.91)	6.66(1.75)	5.02(1.10)	4.17(0.80)
EMMAX(FDR=0.05)	18.91(0.57)	18.85(0.58)	18.54(0.69)	18.14(0.75)
EMMAX(FDR=0.1)	19.28(0.57)	19.24(0.55)	18.94(0.72)	18.59(0.75)
Method	$\tau = 15$	$\tau = 20$	$\tau = 50$	$\tau = 100$
fPOCRE(L1)$_{\gamma=0.1}$	15.49(1.76)	14.65(2.01)	7.44(5.96)	7.43(5.60)
fPOCRE(L1)$_{\gamma=1}$	1.51(0.52)	1.51(0.62)	1.57(0.70)	1.28(0.85)
fPOCRE(EBTZ)$_{\gamma=0.1}$	15.51(1.64)	14.03(2.61)	5.73(4.89)	3.89(3.42)
fPOCRE(EBTZ)$_{\gamma=1}$	1.84(0.42)	1.86(0.58)	1.88(0.74)	1.14(1.00)
MLMM$_{\gamma=0.1}$	4.79(0.95)	4.22(0.95)	3.06(0.89)	2.11(0.83)
MLMM$_{\gamma=1}$	3.76(0.79)	3.42(0.78)	2.04(0.72)	1.31(0.46)
EMMAX(FDR=0.05)	17.80(0.93)	17.25(1.23)	13.37(2.43)	10.46(2.36)
EMMAX(FDR=0.1)	18.18(0.90)	17.83(1.06)	14.29(2.50)	11.04(2.30)

The results of the three linkage groups are shown in Table 2.3 and Table 2.4. Similar to what we observe in the one linkage group simulation study, even though EMMAX has good power but with the cost of greatly increased false positives. On the other hand, MLMM with $\gamma = 1$ has relatively low false detections, but suffers from loss of power. fPOCRE with $\gamma = 0.1$ has both good power and reduced false discoveries

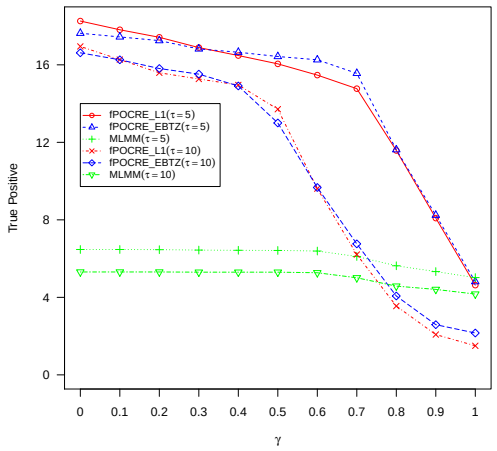
Table 2.4.: Performance comparison in the three linkage groups simulation study. Reported are the mean false positives (FP) across 100 simulated data sets with standard errors presented in the parentheses.

Method	$\tau = 0.5$	$\tau = 1$	$\tau = 5$	$\tau = 10$
fPOCRE(L1)$_{\gamma=0.1}$	3.52(2.24)	3.30(2.03)	1.95(1.71)	1.51(1.46)
fPOCRE(L1)$_{\gamma=1}$	0.70(0.89)	0.57(0.85)	0.04(0.20)	0.00(0.00)
fPOCRE(EBTZ)$_{\gamma=0.1}$	3.01(2.54)	2.36(1.81)	1.86(1.35)	1.98(2.07)
fPOCRE(EBTZ)$_{\gamma=1}$	1.05(0.93)	0.78(0.83)	0.08(0.31)	0.00(0.00)
MLMM$_{\gamma=0.1}$	11.42(5.86)	13.62(6.15)	20.07(4.07)	22.23(3.72)
MLMM$_{\gamma=1}$	1.67(1.21)	1.66(1.22)	1.15(0.85)	0.74(0.68)
EMMAX(FDR=0.05)	35.39(3.65)	34.52(3.89)	29.46(5.02)	24.56(5.22)
EMMAX(FDR=0.1)	46.17(6.00)	45.59(6.32)	39.15(6.35)	33.45(7.04)
Method	$\tau = 15$	$\tau = 20$	$\tau = 50$	$\tau = 100$
fPOCRE(L1)$_{\gamma=0.1}$	1.54(1.58)	1.45(1.61)	2.04(6.80)	13.62(16.73)
fPOCRE(L1)$_{\gamma=1}$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
fPOCRE(EBTZ)$_{\gamma=0.1}$	2.14(1.96)	1.70(1.81)	0.47(1.31)	2.67(9.36)
fPOCRE(EBTZ)$_{\gamma=1}$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
MLMM$_{\gamma=0.1}$	23.67(2.57)	24.24(2.16)	25.84(1.24)	26.81(0.95)
MLMM$_{\gamma=1}$	0.52(0.69)	0.38(0.58)	0.27(0.53)	0.08(0.27)
EMMAX(FDR=0.05)	20.92(5.33)	17.96(5.50)	7.31(4.04)	2.73(2.75)
EMMAX(FDR=0.1)	28.41(7.43)	24.46(6.98)	11.16(5.76)	4.91(5.31)

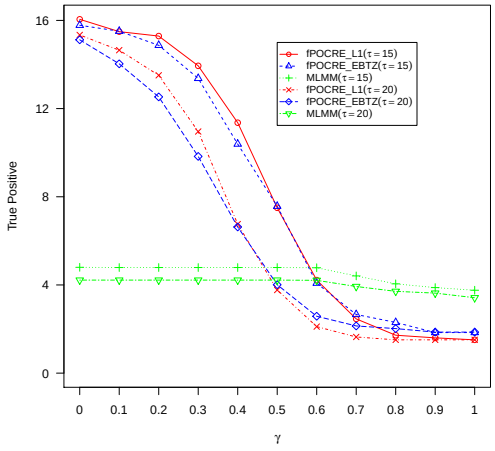
in all the different τ settings considered. Again in general, fPOCRE$(L_1)_{\gamma=0.1}$ gives slightly higher true positives over fPOCRE(EBTZ)$_{\gamma=0.1}$.



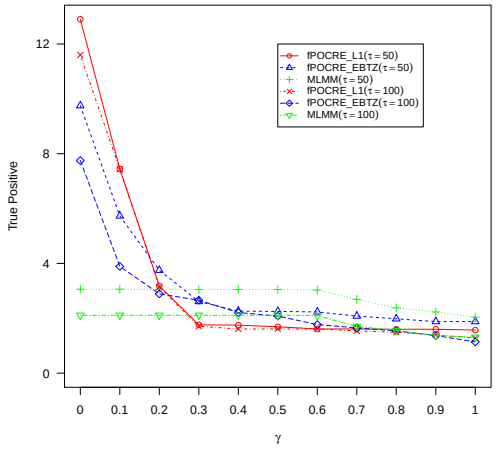
(a) True positives: $\tau = 0.5$ & $\tau = 1$



(b) True positives: $\tau = 5$ & $\tau = 10$

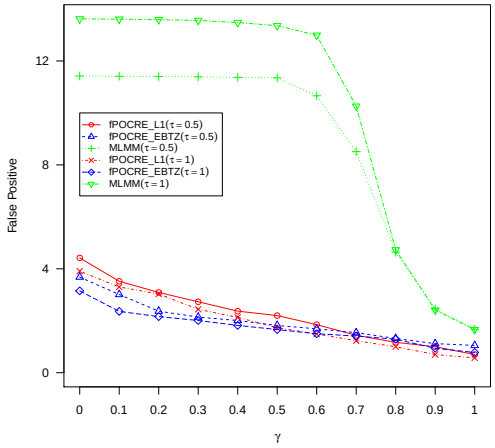


(c) True positives: $\tau = 15$ & $\tau = 20$

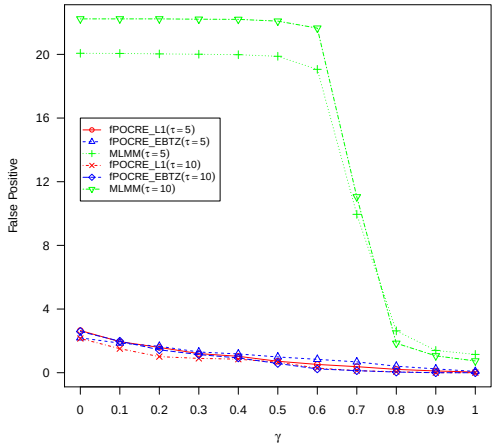


(d) True positives: $\tau = 50$ & $\tau = 100$

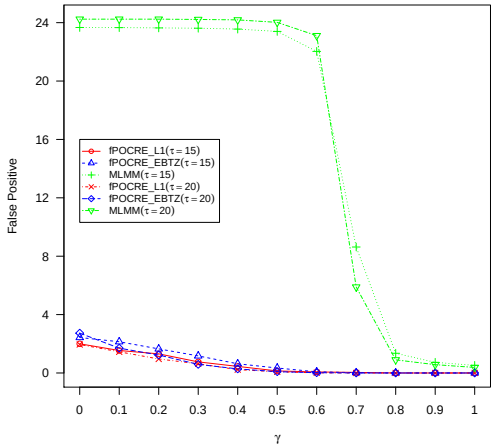
Figure 2.3.: True positives of fPOCRE(L_1), fPOCRE(EBTZ) and MLMM algorithms in the simulation study of three linkage groups



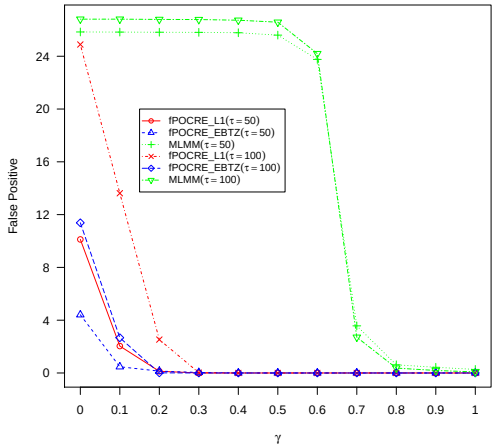
(a) False positives: $\tau = 0.5$ & $\tau = 1$



(b) False positives: $\tau = 5$ & $\tau = 10$



(c) False positives: $\tau = 15$ & $\tau = 20$



(d) False positives: $\tau = 50$ & $\tau = 100$

Figure 2.4.: False positives of fPOCRE(L_1), fPOCRE(EBTZ) and MLMM algorithms in the simulation study of three linkage groups

Figure 2.3 and Figure 2.4 compare the mean true positives and false positives of fPOCRE and MLMM with γ increases. Similar to the conclusion we draw from the first simulation study, fPOCRE outperforms MLMM in terms of having a larger number of true positives and a lower number of false positives. When *gamma* is smaller, the difference between fPOCRE and MLMM is bigger.

2.5 Real Data Analyses

2.5.1 Application to the Regional Mapping Panel *A. Thaliana* Data Set

Horton *et al.* 2012 studied high-resolution description of the global pattern of genetic variation using worldwide *Arabidopsis thaliana* accessions from the Regional Mapping panel. There are 336 *A. thaliana* samples measured with sodium accumulation, and genotyped at 200155 SNPs. We apply both fPOCRE using L_1 penalty and MLMM algorithms to this data set. When applying MLMM algorithm, we notice that a pre-determined step-size needs to be given before applying MLMM, and there is no good thumb of rule to choose prior to this selection, we use step size 10 in this Arabidopsis sodium level analysis. The results are presented in Table 2.5.

fPOCRE$_{\gamma=0.1}$ identifies six signals while MLMM$_{\gamma=1}$ finds four SNPs. The SNP (chromosome 4: 6,392,280) resides in the first exon of gene *AtHKT1;1* that is previously reported for association with sodium accumulation is identified by both fPOCRE and MLMM methods. fPOCRE$_{\gamma=0.1}$ selects three more SNPs that are near the gene *AtHKT1;1* and they are all within 50kb of the SNP (chromosome 4: 6,392,280), while MLMM$_{\gamma=1}$ selects two more SNPs that are nearby. Furthermore, fPOCRE is much more computational efficient, it takes about one minute to finish analyzing this data, MLMM on the other hand uses about 13 minutes. The computational improvement will be more obvious with a much larger data set.

In hypothesis testing, p-value is a popular indicator to quantify statistical significance. We then try to assign a p-value to the SNPs we find. Recently, Meinshausen *et al.* (2009) has proposed a multi-split method to assign statistical significance and

Table 2.5.: Arabidopsis real data analysis results. Reported are the coefficient estimates, with p-values in the parenthesis.

Chr	Physical Position	fPOCRE$_{\gamma=0.1}$(p-value*)	MLMM$_{\gamma=1}$(p-value**)
2	12471135	26.18(1.33E-02)	-
3	7966725	83.71(2.99E-02)	-
3	21670298	-	252.20(1.68E-07)
4	6388940	235.63(2.27E-02)	-
4	6392280	485.27(4.36E-09)	616.44(3.93E-27)
4	6418442	82.52(1.48E-07)	272.97(3.82E-15)
4	6719618	326.62(2.89E-02)	323.61(3.03E-08)

*p-values are calculated using the 10-split method

**p-values are calculated from the final model with selected SNPs

construct p-values for high-dimensional analyses where the number of predictors may be much larger than the sample size. In each split, the data is divided into two parts, fPOCRE uses the first part and builds a statistical model, then a classical variable selection technique is applied to the selected variables using the data from the second part. The method has the property of asymptotic error control and model selection consistency. Here we apply multi-split method with a total of 10 splitting. Reported are the ones with a p-value less than 0.05.

2.5.2 Application to the Heterogeneous Mice Data

Mouse is an important model organism for understanding gene functions in mammals and population structure would be expected in data sets with heterogeneous mice. Legarra *et al.* 2008; Valdar *et al.* 2006 performed a genome-wide association study using heterogeneous mice data that generated from eight inbred lines. A total of 1872 mice with pedigree information are genotyped at 11730 SNPs. We apply both

fPOCRE with L_1 penalty and MLMM with step size 10 to study the potential causal SNPs of the weight growth slope.

After the analysis, two SNPs on chromosome 18 (18:57650519, rs3705107 and 18:63716343, rs3023468) are chosen by fPOCRE$_{\gamma=0.1}$ model, where the SNP rs3705107 is near the previously identified 95% confidence interval of the QTL (18:55704779-57510467, Valdar et al. 2006) influencing the weight growth slope. On the other hand, nothing is selected by MLMM$_{\gamma=1}$. Detailed results could be found in Table 2.6. P-values associated with fPOCRE results are calculated through the multi-splitting method with 10 splits.

Table 2.6.: fPOCRE results of heterogeneous mouse data. Reported are the coefficient estimates, with p-values in the parenthesis

Chromosome	SNP	Physical Position	Genetic Map	Beta(p-value*)
18	rs3705107	57650519	36.85	3.61E-03(3.88E-02)
18	rs3023468	63716343	44.56	4.24E-03(2.75E-02)

*p-values are calculated using the 10-split method

2.6 Conclusion

In summary, we have presented an efficient feature selection method for family data in the large p small n scenario. Our proposed approach is a hybrid of the penalization method and the linear mixed model and is computationally tractable for moderately large data set which is quite common due to nowadays high throughput technology. The fPOCRE algorithm works by simultaneously incorporating tens of thousands of genetic markers in a single statistical model, building orthogonal components where a penalty is included to force most regression coefficient to be exactly zero. Being a multi-locus algorithm separates us from most of other algorithms that are also in the linear mixed model framework. Our new analytical procedure has more

power and the potential to add a significant number of discoveries in a genome-wide association study. It also appears to be less open to false positives than the single-locus based methods EMMAX when a non-zero correlation structure exists between associated markers and the underlying genetic architecture is polygenic. MLMM, a stepwise linear mixed model regression which is also a multi-SNP algorithm has been shown to perform well in GWAS. However, it suffers from instability of stepwise regression and loss of power when variables are correlated.

Our simulation study demonstrates fPOCRE has promising performance over both EMMAX and MLMM in terms of improved power and reduced false positives when the causal effects are from structurally correlated SNPs. Simulations based on the real genetic structure with different genetic variation to residual variation ratios are used to compare the performance of different methods. In general, fPOCRE with L_1 penalty would be suggested for its computational efficiency. The final model is chosen by eBIC, and a larger tuning parameter (γ) is recommended if the false discovery rate is more concerned, otherwise, smaller values can increase power in order to detect weaker signals.

An *Arabidopsis* and a heterogeneous mouse data have been used to demonstrate the advantages of using fPOCRE. A reported association of the *Arabidopsis* sodium accumulation is confirmed in both fPOCRE and MLMM results. The optimal fPOCRE model includes three additional SNPs that are within the associated region. On the other hand, MLMM finds two more genetic markers in the same genetic region. In our mouse data analysis, one of our two findings locates in a previous reported association region, however, MLMM finds none. Moreover, our proposed algorithm fPOCRE is computational efficient and runs much faster than the MLMM in all our real data analyses.

The proposed hybrid algorithm provides a clear alternative way to perform a family-based multi-SNP GWA study and more extensions could be further investigated under this framework, such as case-control GWAS and multiple traits in structured associations.

3. AN EFFICIENT METHOD FOR CASE-CONTROL GENOME-WIDE ASSOCIATIONS WITH FAMILY STRUCTURE

In this chapter, we describe the details of a proposed efficient method for case-control GWAS with family structure. Our proposed algorithm GEE-POCRE constructs penalized orthogonal components under the framework of generalized estimating equation (GEE). Moreover, it simultaneously selects variables when assuming correlated structures among study samples. We compare GEE-POCRE with gPOCRE, an algorithm that is multi-locus regression but does not consider family structure, and a set of algorithms that take account of the family structure but are single-locus methods, such as TDT, FBAT, ROADTRIPS. Our simulation studies demonstrate GEE-POCRE has promising performance over its competitor algorithms in terms of both power and false positives. A real data analysis is also included to show the performance of GEE-POCRE.

We start this chapter with the motivation of our GEE-POCRE algorithm. Section 3.2 describes the GEE model, the penalized orthogonal components regression for the generalized linear model (gPOCRE) and how to further extend penalized orthogonal components regression for the GEE model. The full details on the algorithm are also provided in this section. Section 3.3 and Section 3.4 contain the results of simulations and a real data analysis respectively. We conclude this chapter with a final discussion.

3.1 Introduction and Motivation

Genome-wide association studies have been frequently conducted to help identify genetic loci associated with complex traits and human diseases, such as schizophrenia, Late Onset Alzheimer's Disease and type II diabetes. It is well-known that spurious associations may occur if hidden population structure and cryptic relatedness are

not properly accounted. For different study designs and different types of traits, many statistical approaches have been developed to correct for population structure and cryptic relatedness to identify hidden genetic risk factors. Case-control is a popular study design in epidemiology and pharmacology.For certain types of study designs, commonly used are family-based testing methods including the transmission disequilibrium test (TDT), and family-based association Test (FBAT). They test for the differences between the observed offspring genotypes and the expected ones under Mendelian rule. Both these methods are linkage analyses, and are virtually immune to confounding. However, family-based tests are in general less powerful than case-control association methods because they require more samples to obtain good power. Moreover, such family-based data is more than likely not obtainable in many situations.

While methods based on linear mixed models have shown advantages in analyzing GWAS with normal traits but how to extend them to case-control GWAS is not clear yet. Here we will establish a GEE model for case-control GWAS, and develop a supervised dimension reduction method to identify important SNPs.

As shown in Liang and Zeger (1986) that GEE provides a framework to analyze data sets with correlated observations in longitudinal study. Instead of specifying the joint distribution, GEE assumes a marginal mean and covariance model that uses a user-defined working correlation matrix (Chen *et al.* 2011). Moreover, GEE is known to be robust to mis-specification of the working correlation matrix and has shown successes in many studies. In family-based GWAS, subjects can be grouped within each family. The sample relatedness can be modeled through the kinship matrix which is included in the working correlation matrix.

With nowadays high-throughput technology, numerous genetic markers are genotyped at a relatively low price, making it challenge to statistically analyze such large p small n data. Dimension reduction is an important method that helps reduce the dimensionality of the variable space before fitting any models. The partial least squares (PLS) method proposed by Wold (1975) is considered one of the commonly used sta-

tistical strategies to capture data information in a lower dimension space. Later, Lin *et al.* (2014) propose the penalized orthogonal components regression in generalized linear models (gPOCRE) that extends PLS to fit high dimensional generalized linear models, it does the job of variable selection and estimation simultaneously.

In this paper we examine the use of GEE for family-based case-control GWAS, and follow gPOCRE to develop a variable selection method, aka GEE-POCRE, to identify genetic variants of binary traits. The proposed algorithm assumes a model with multiple genetic markers, making it a multi-locus approach compared to TDT and FBAT, which are essentially single-locus methods. Gained power would be expected by using GEE-POCRE if the disease status is truly influenced by multiple genetic factors.

3.2 Methods

3.2.1 Generalized Estimating Equations

Generalized estimating equations (GEE), first introduced by Liang and Zeger (1986), has become very popular in epidemiology, pharmacology and other related research areas. It extends the generalized linear model (GLM) to handle correlated observations through a user-defined working correlation matrix and further assumes a general mean-covariance structure (Zeger and Liang , 1986).

Suppose correlated study subjects are put into the same cluster. Let $Y_i = (y_{i1}, y_{i2}, ..., y_{in_i})^T$ be a vector of outcomes from cluster i, $\mathbf{X}_i = (X_{i1}, ..., X_{in_i})^T$ be the $n_i \times p$ design matrix for the ith cluster, $i = 1, ..., K$. We assume that the marginal density of y_{ij} is

$$f(y_{ij}) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right], \tag{3.1}$$

where $a(\phi)$ is a function of the dispersion parameter ϕ, and $\theta_{ij} = g(\eta_{ij})$ with

$$\eta_{ij} = X_{ij}\beta.$$

It leads to the general mean and covariance structure of the exponential families

$$\mu_{ij} = E(y_{ij}) = b'(\theta_{ij}), \tag{3.2}$$

$$a_{ij} = var(y_{ij}) = b''(\theta_{ij})a(\phi). \tag{3.3}$$

GEE differs from GLM in that it has a user-specified covariance structure that takes correlations among observations into account to increase efficiency. Let $\mathbf{R}(\tau)$ be an $n \times n$ symmetric working correlation matrix with an unknown parameter τ, then the covariance matrix can be defined as

$$\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}(\tau)\mathbf{A}_i^{1/2}, \tag{3.4}$$

where $\mathbf{A}_i = \text{diag}(a_{ij})$.

Then the generalized estimating equation is expressed as

$$\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i{}^{-1}(Y_i - b_i'(\theta)) = 0, \tag{3.5}$$

where $\mathbf{D}_i = \dfrac{\partial b_i'(\theta)}{\partial \beta} = \mathbf{A}_i\mathbf{\Delta}_i\mathbf{X}_i$, $\mathbf{\Delta}_i = \text{diag}(\partial\theta_{ij}/\partial\eta_{ij})$.

There is no much difference between GEE and GLM except that the variance covariance matrix $\mathbf{V}$ is no longer a diagonal matrix, with non-zero values on the off-diagonal.

The following iterative re-weighted least square (IRWLS) algorithm can be employed to compute β,

$$\beta^{(t+1)} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}Z, \tag{3.6}$$

where the working response Z and the weight matrix $\mathbf{W}$ are defined as

$$Z = \eta + \frac{\partial \eta}{\partial \mu} \times (Y - \mu), \tag{3.7}$$

$$\mathbf{W} = var(Y)^{-1} \times \left(\frac{\partial \mu}{\partial \eta}\right)^2. \tag{3.8}$$

However, the way to estimate β is no longer applicable when dealing with large p small n data. An alternative approach needs to be considered instead.

3.2.2 Penalized Orthogonal Components Regression in the Generalized Linear Model

As suggested by Lin *et al.* (2014) that gPOCRE intends to build up a sequence of orthogonal components in high-dimensional GLMs and subsequently put a penalization function on the coefficient of the predictors to identify sparse signals. The bias correction by Firth (1993) is also applied. Now assuming the response Y is a member of the exponential family distribution and its density function is expressed as

$$f(y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right], \tag{3.9}$$

where θ and ϕ are the canonical parameter and the dispersion parameter respectively. The mean $\mu = E(Y|X)$ is related to the predictors through a link function $g(.)$,

$$g(\mu) = \mu_0 + \mathbf{X}\beta. \tag{3.10}$$

Here μ_0 is the intercept, $\mathbf{X}$ is a design matrix, β is a p-dimensional column vector containing all the regression coefficients of the predictors. The orthogonal components are sequentially constructed with a pre-specified weight $\mathbf{W}$ and the design matrix $\mathbf{X}$ is column-wise centralized e.g. $E(W\mathbf{X}) = 0$ prior to the construction. Starting with $\mathbf{X}_1 = \mathbf{X}$ and assuming the first j orthogonal components have been constructed, we now proceed to obtain the $(j+1)$-st orthogonal component. $\mathbf{X}_{j+1}$ is defined as $\mathbf{X}_{j+1} = \mathbf{X}_j - \mathbf{X}_j\alpha_j\theta_j$ and it is orthogonal to $\mathbf{X}_j\alpha_j$. With η estimated from the previous step, say η_j, the working response $Z(\eta)$ is calculated as

$$Z(\eta) = \eta + \frac{\partial\eta}{\partial\mu} \times (Y - \mu), \tag{3.11}$$

where $\mu = g^{-1}(\eta)$, then the loadings of the $(j+1)$-st component is updated with

$$\alpha(\eta) = E(\mathbf{X}_{j+1}^T\mathbf{W}Z(\eta))/||E(\mathbf{X}_{j+1}^T\mathbf{W}Z(\eta))||. \tag{3.12}$$

Then update $\eta(\alpha)$ as

$$\eta(\alpha) = E(\mathbf{W}Z)/E(\mathbf{W}) + \sum_{k=1}^{j}\mathbf{X}_k\alpha_k\nu_k + \mathbf{X}_{j+1}\alpha\nu, \tag{3.13}$$

where $\nu = E(\alpha^T \mathbf{X}_{j+1}^T \mathbf{W} Z)/E(\alpha^T \mathbf{X}_{j+1}^T \mathbf{W} \mathbf{X}_{j+1} \alpha)$ and $\nu_k = E(\alpha_k^T \mathbf{X}_j^T \mathbf{W} Z)/E(\alpha_k^T \mathbf{X}_j^T \mathbf{W} \mathbf{X}_j \alpha_k)$ for $k = 1, ..., j$. We can iterate between $\alpha(\eta)$ and $\eta(\alpha)$ until $\alpha(\eta)$ converges.

Noticing that each orthogonal component $\mathbf{X}_j \alpha_j$ could be re-expressed as $\mathbf{X}\varpi_j = \mathbf{X} \prod_{l=1}^{j-1} (I - \alpha_{j-l}\theta_{j-l})\alpha_j$. The fitted generalized orthogonal component regression can be written as

$$g(\mu) = \mu + \sum_j \mathbf{X}\varpi_j \vartheta_j. \tag{3.14}$$

To enforce sparsity, a penalty function is added when constructing the sequence of orthogonal components. Cross-validation is then explored to help select the tuning parameter of the penalty function.

3.2.3 GEE Models for Family-Based Case-Control GWAS

Here we propose a GEE model for the case-control GWAS. specifically, we assume the following model,

$$g(\mu) = \eta = \mu_0 + \mathbf{X}\beta = \mu_0 + \sum_j (\mathbf{X}\varpi_j)\vartheta_j. \tag{3.15}$$

The variance-covariance matrix $\mathbf{V}$ is formulated as

$$\mathbf{V} = \mathbf{A}^{1/2} \mathbf{R} \mathbf{A}^{1/2}, \tag{3.16}$$

where

$$\mathbf{A} = \text{diag}(a_1, ...a_n), a_i = \text{var}(y_i). \tag{3.17}$$

Motivated by the kinship matrix that captures the sample relatedness used with linear mixed model, we adopt this idea into the GEE model and suggest the following working correlation matrix,

$$\mathbf{R} = (1 - \tau)\mathbf{I} + \tau \mathbf{K}, \tau \in [0, 1]. \tag{3.18}$$

To ease the computation, we first eigen-decompose the kinship matrix $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^T$. With pre-specified τ and $\mathbf{A}$, the weight matrix $\mathbf{W}$ is calculated as

$$\mathbf{W} = \mathbf{V}^{-1} = \mathbf{A}^{-1/2} \mathbf{U} \left\{(1 - \tau)\mathbf{I} + \tau \mathbf{D}\right\}^{-1} \mathbf{U}^T \mathbf{A}^{-1/2}, \tag{3.19}$$

where the inverse of $(1-\tau)\mathbf{I}+\tau\mathbf{D}$ can be easily calculated as D is diagonal.

For given τ and $\mathbf{A}$, we can employ gPOCRE described in the previous section to this GEE model to estimate all ϖ_j and ϑ_j so as to estimate β. Indeed, all the orthogonal components can be sequentially constructed following equations (3.11)-(3.13), and η is then calculated as in (3.15).

A newly estimated β would suggest an update to both τ and $\mathbf{A}$. We can update $\mathbf{A}$ as

$$\mathbf{A} = b''(\theta)/(\bigtriangledown g^{-1}(\eta))^2. \tag{3.20}$$

A moment estimator of τ is sought in our proposed algorithm. Specifically, we will utilize the Pearson's residual, i.e.,

$$e_i = (y_i - \mu_i)/\sqrt{var(y_i)}. \tag{3.21}$$

Since $\rho_{ij} = E(e_i e_j) = corr(y_i, y_j)$ which is the ij-th component in the true correlation matrix, $e_i e_j$ is an unbias estimator of ij-th element of the working correlation matrix $\mathbf{R}$. Therefore, finding τ is equivalent to the following regression problem with respect to the upper off-diagonal elements, e.g., $i < j$,

$$\tau(K_{ij} - I_{ij}) = R_{ij} - I_{ij}, \quad i < j. \tag{3.22}$$

Iterating between construction of penalized orthogonal components and updating τ and $\mathbf{A}$ leads to a fit to the GEE model with a sparse estimate of β.

3.2.4 The Algorithm

In this section, we present the complete GEE-POCRE algorithm which encompasses building the penalized orthogonal components as well as updating the parameters, including τ in the working correlation matrix and the matrix $\mathbf{A}$ with variance on the diagonal.

Following the idea described in the previous section, let the outcome variable and the design matrix denoted by $Y = (y_1, ..., y_n)^T$ and $\mathbf{X} = (X_1, ..., X_n)^T$ respectively. For each fixed tuning parameter λ, we can proceed the algorithm as follows.

0. Let initial values $\tau^{(0)} = 0$ and $\mathbf{A}^{(0)} = n$, so that $\mathbf{A}^{(-1/2)} = 1/\sqrt{n}$.

1. Given τ and $\mathbf{A}$, we update the weight matrix $\mathbf{W} = \mathbf{A}^{(-1/2)}\mathbf{U}\{(1-\tau\mathbf{I}) + \tau\mathbf{D}\}^{-1}\mathbf{U}^T\mathbf{A}^{(-1/2)}$, and centralize $\mathbf{X}$ appropriately according to $\mathbf{W}$, e.g.$\mathbf{X}^T\mathbf{W}1_n = 0_p$. We then obtain estimate of β via constructed penalized orthogonal components, i.e.,

$$\hat{\beta} = \sum_j \zeta_j \alpha_j \gamma_j, \tag{3.23}$$

where

$$\begin{cases} \zeta_j = \prod_{k=1}^{j-1}(\mathbf{I}_p - \alpha_k \mathbf{P}_k), \\ \gamma_j = \alpha_j^T \mathbf{X}_j^T \mathbf{W}\mathbf{Z}/\alpha_j^T \mathbf{X}_j^T \mathbf{W}\mathbf{X}_j\alpha_j, \\ \mathbf{P}_j = \alpha_j^T \mathbf{X}_j^T \mathbf{W}\mathbf{X}_j/\alpha_j^T \mathbf{X}_j^T \mathbf{W}\mathbf{X}_j\alpha_j. \end{cases}$$

Here the loadings α_j are obtained through the following iterative procedure, starting with $\eta_1 = \log\left(\frac{\bar{Y}}{1-\bar{Y}}\right)$, $\mathbf{X}_1 = \mathbf{X}$, and $j = 1$.

1.a. Let $\eta_j = \eta_{j-1}$;

1.b. Update $Z = \eta_j + H^{-1}(Y - g^{-1}(\eta_j))$, with $H = \mathrm{diag}(\bigtriangledown g^{-1}(\eta_{j1}), ..., \bigtriangledown g^{-1}(\eta_{jn}))$;

1.c. Update $\mu_0 = 1_n^T\mathbf{W}Z/(1_n^T\mathbf{W}1_n)$;

1.d. Calculate $(\nu, \xi) = \arg\min_{\nu,\xi:\|\xi\|=1}\left\{-2\nu^T\mathbf{X}_j^T Z\mathbf{W}Z^T\mathbf{X}_j\xi + \|\nu\|^2 + P_\lambda(\nu)\right\}$, and update $\alpha_j = \nu/\|\nu\|$;

1.e. Update $\gamma_k = \alpha_k^T\mathbf{X}_k^T\mathbf{W}Z/(\alpha_k^T\mathbf{X}_k^T\mathbf{W}\mathbf{X}_k\alpha_k)$, $k = 1, ..., j$;

1.f. Update $\eta_j = \mu_0 1_n + \sum_{k=1}^{j}\mathbf{X}_k\alpha_k\gamma_k$.

1.g. Iterate between 1.b. and 1.f. until α_j converges;

1.h. This whole procedure stops if $\alpha_j = 0$. Otherwise, calculate $\mathbf{P}_j = \alpha_j^T\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j/(\alpha_j^T\mathbf{X}_j^T\mathbf{W}\mathbf{X}_j\alpha_j)$ and $\mathbf{X}_{j+1} = \mathbf{X}_j\alpha_j\mathbf{P}_j$, then start over at 1.a. with $j = j + 1$.

2. If there is at least one component constructed from the previous step, update both τ and $\mathbf{A}$ as follows. The algorithm stops otherwise.

2.a. Calculate $\mathbf{A} = b''(\theta)/(\bigtriangledown g^{-1}(\eta))^2$, where $b''(\theta) = \pi(1-\pi)$, and $\pi = g^{-1}(\eta)$;

2.b. Calculate $\tau = \sum_{i<j}(R_{ij}K_{ij})/\sum_{i<j}K_{ij}^2$, where $R_{ij} = e_ie_j$ with $e_i = (Y - b'(\theta))/\sqrt{A}$.

3. Repeat 1 with updated τ and $\mathbf{A}$, and therefore obtain final estimate of β.

Here 10-fold cross validation is applied to choose the optimal tuning parameter λ in GEE-POCRE.

3.3 Simulation Study

For our simulation study, we will use the genotypic values of an existing Japanese Schizophrenia study (Yamada *et al. 2011*). This study includes a total of 120 trio-families, with 4616 SNPs genotyped on chromosome 4. We simulate phenotypic values by assuming causal genetic variants only existing on chromosome 4, and maintaining the trio-family structure.

The binary trait Y follows a logistic model

$$p = E(Y) = \frac{\exp(\mu + \mathbf{X}\beta)}{1 + \exp(\mu + \mathbf{X}\beta)}. \tag{3.24}$$

Let p_1 and p_2 denote the marginal probabilities of the parents being case (i.e., $Y = 1$). The phenotypic values of the parents can be simulated through

$$Y_1 \sim \text{Bernoulli}(p_1), \tag{3.25}$$

$$Y_2 \sim \text{Bernoulli}(p_2). \tag{3.26}$$

Let $p(Y_1, Y_2)$ be the probability of the offspring being case conditional on the case-control status of both parents. Then,

$$Y_3|Y_1, Y_2 \sim \text{Bernoulli}(p(Y_1, Y_2)). \tag{3.27}$$

Now let ρ denote the correlation coefficient between the phenotypic values of parents and the offspring, and p_3 denote the marginal probability of the offspring. We then have the following equations derived from that $corr(Y_1, Y_3) = corr(Y_2, Y_3) = \rho$ and $E[Y_3] = E[p(Y_1, Y_2)] = p_3$.

$$(1 - p_2)p(1, 0) + p_2 p(1, 1) - p_3 - \rho\sqrt{((1 - p_1)p_3(1 - p_3)/p_1)} = 0, \tag{3.28}$$

$$(1 - p_1)p(0, 1) + p_1 p(1, 1) - p_3 - \rho\sqrt{((1 - p_2)p_3(1 - p_3)/p_2)} = 0, \tag{3.29}$$

$$(1 - p_2)p(0, 0) + p_2 p(0, 1) - p_3 - \rho\sqrt{(p_1 p_3(1 - p_3)/(1 - p_1))} = 0. \tag{3.30}$$

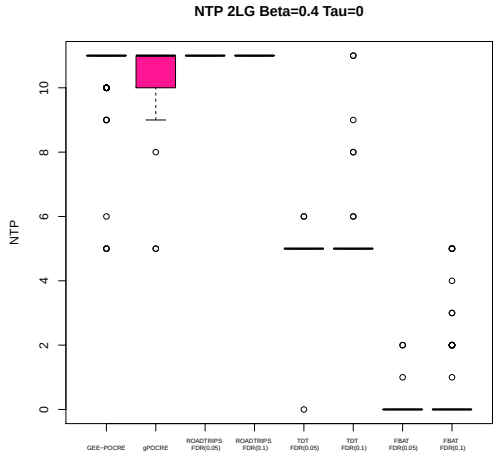
Assuming the parents are exchangeable, that is $p(0,1) = p(1,0)$, we solve the equations above and simulate the offspring phenotypic values based on the conditional probabilities.

Here we present two simulation scenarios with signals clustered in two linkage groups. We evaluate our GEE-POCRE along with some existing methods in terms of power and false discoveries. The methods we compare with are TDT, FBAT, and ROADTRIPs. All of these three are single-locus methods for family-based GWAS. We also apply gPOCRE, a multi-locus method that does not use the pedigree information, to the simulated data.

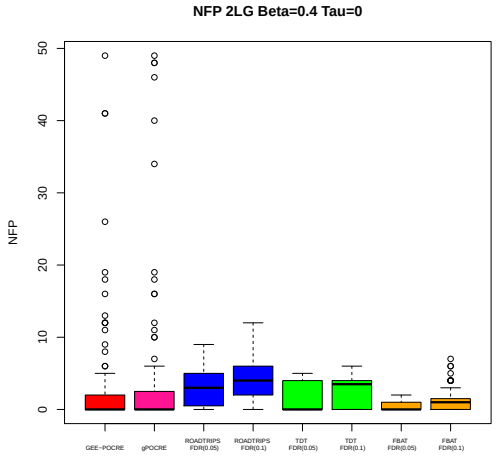
Case I. Eleven Causal SNPs Existing in Two Linkage Groups

Denote the eleven causal SNPs as $X_1, \cdots, X_{11}$. The underlying logistic model is $\text{logit}(E(Y)) = \mu + 0.4\sum_{j=1}^{11} X_j$, where $\text{var}(Y) = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$, A is a diagonal matrix with the variances of Y as the diagonal elements, and $\mathbf{R}$ is a correlation matrix defined as $\mathbf{R} = (1-\tau)\mathbf{I} + \tau\mathbf{K}$. We consider different values of τ, ranging from 0 to 1 with increment 0.1. For each τ, we simulate 100 data sets. Two sets of results are reported with ROADTRIPS, TDT and FBAT by controlling FDR at 0.05 and 0.1 respectively.

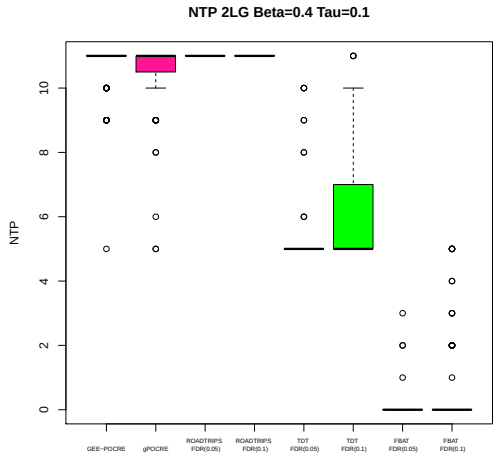
The boxplots of the true positives and false positives among all the compared methods are shown in Figure 3.1. Across all the different τ, GEE-POCRE, gPOCRE, and ROADTRIPS, these three methods have the highest median true positives. On the other hand, GEE-POCRE, TDT and FBAT demonstrated their strength in controlling false discoveries. Apparently, GEE-POCRE has both improved power and a better control of false positives compared with other four methods. With increased τ, the variance of false discoveries of GEE-POCRE and gPOCRE increases, indicating a decreased performance of these two methods. Figure 3.2 summarizes the median true positives and false positives of all compared methods at different τ. We observe a similar trend that GEE-POCRE outperforms all other methods, with a slightly lower number of false positives when τ is small.



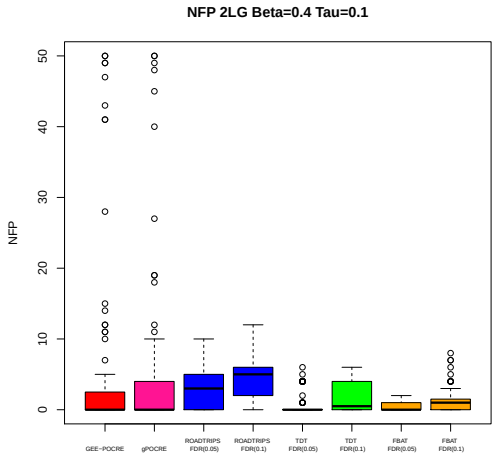
(a) $\tau = 0$: True Positives



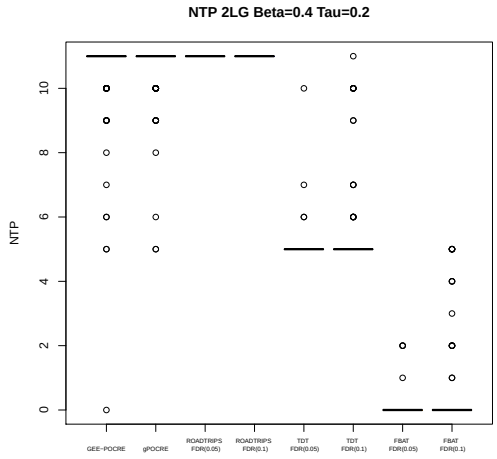
(b) $\tau = 0$: False Positives



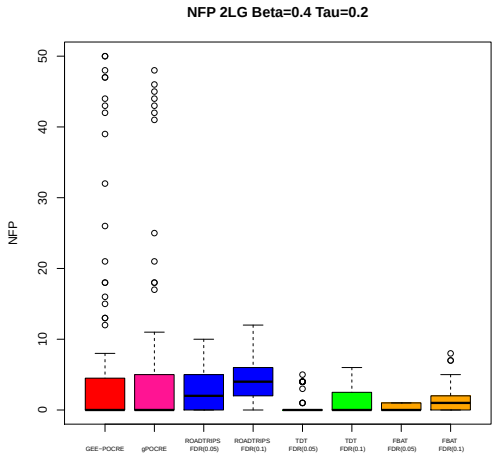
(c) $\tau = 0.1$: True Positives



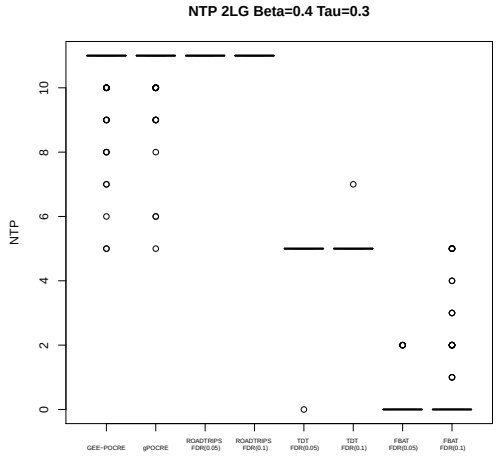
(d) $\tau = 0.1$: False Positives



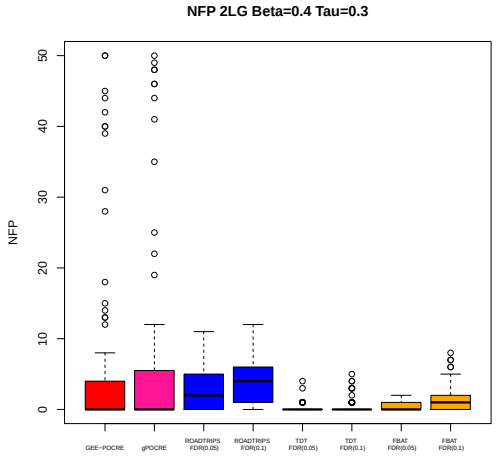
(e) $\tau = 0.2$: True Positives



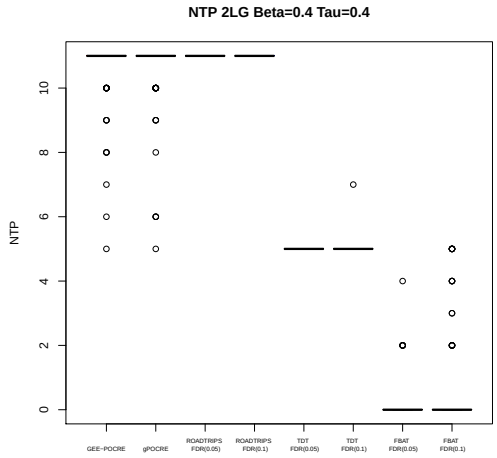
(f) $\tau = 0.2$: False Positives



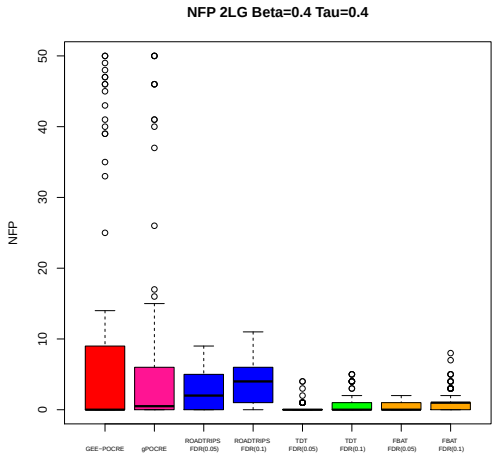
(a) $\tau = 0.3$: True Positives



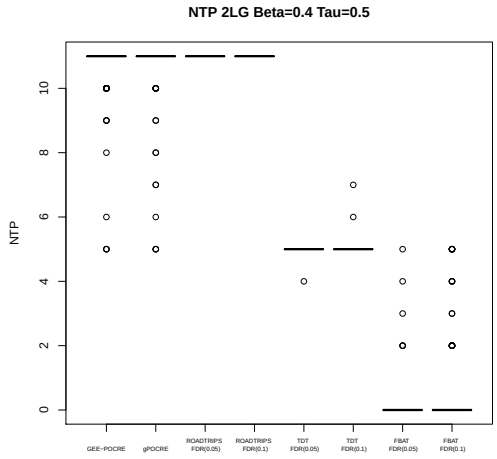
(b) $\tau = 0.3$: False Positives



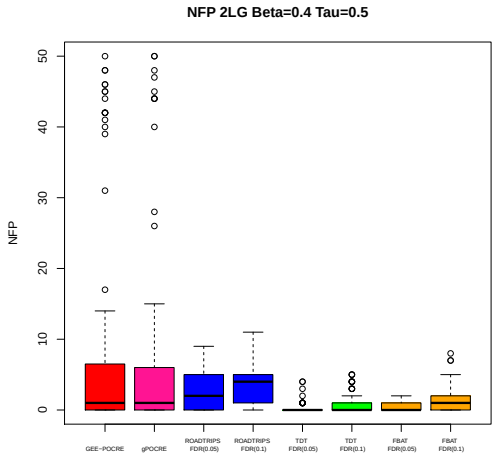
(c) $\tau = 0.4$: True Positives



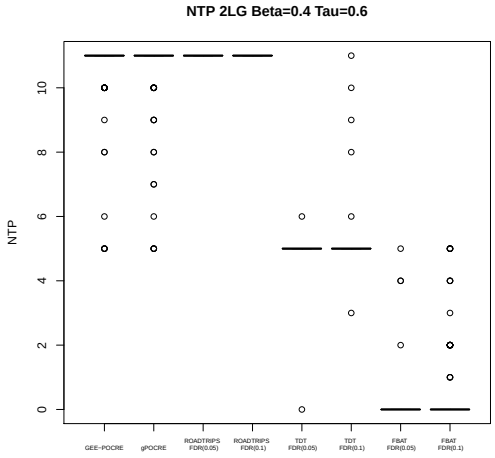
(d) $\tau = 0.4$: False Positives



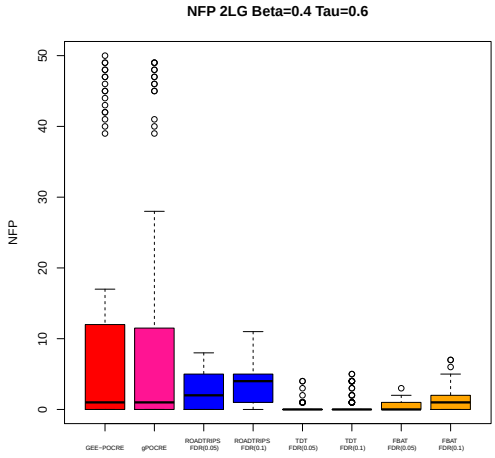
(e) $\tau = 0.5$: True Positives



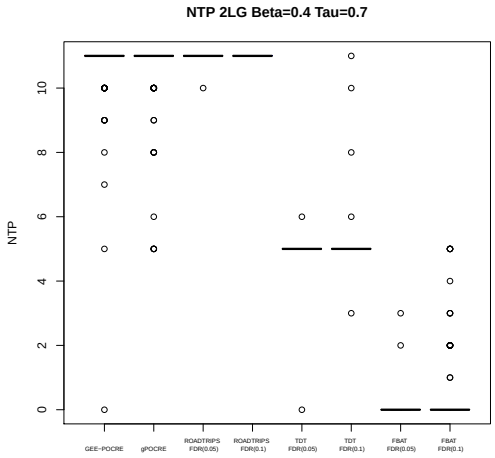
(f) $\tau = 0.5$: False Positives



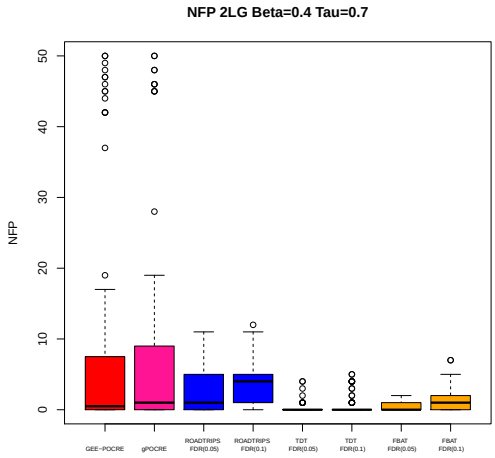
(a) $\tau = 0.6$: True Positives



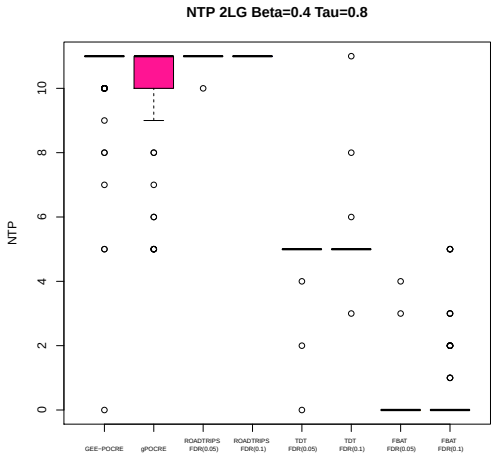
(b) $\tau = 0.6$: False Positives



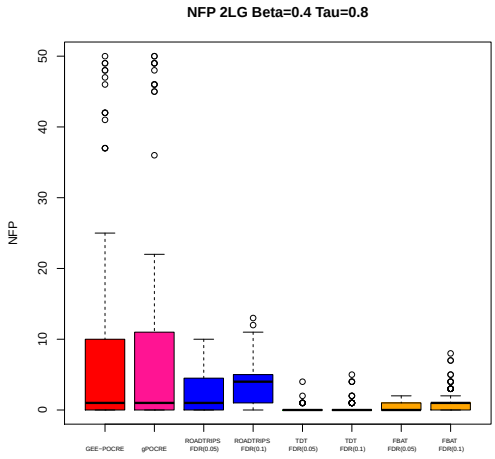
(c) $\tau = 0.7$: True Positives



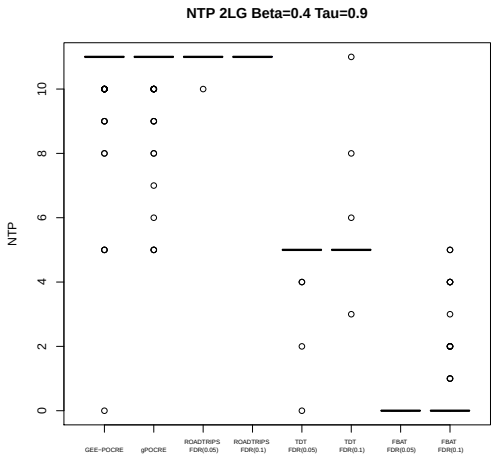
(d) $\tau = 0.7$: False Positives



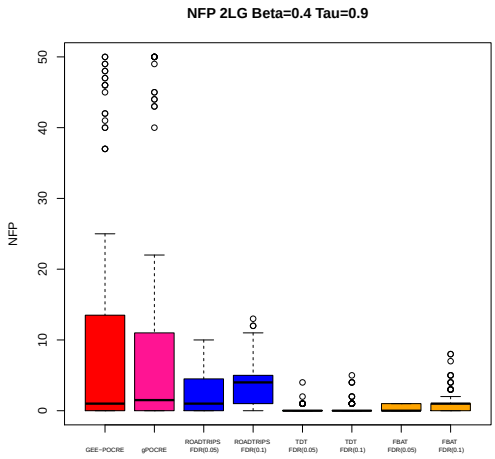
(e) $\tau = 0.8$: True Positives



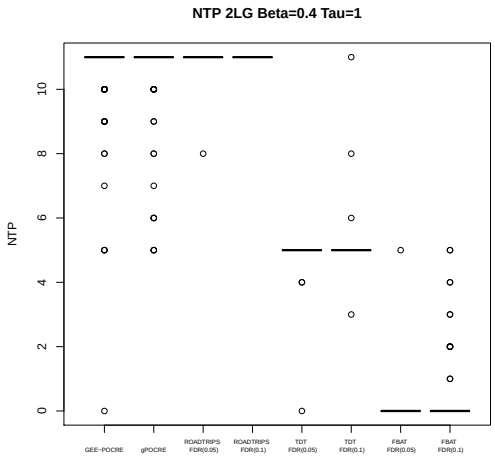
(f) $\tau = 0.8$: False Positives



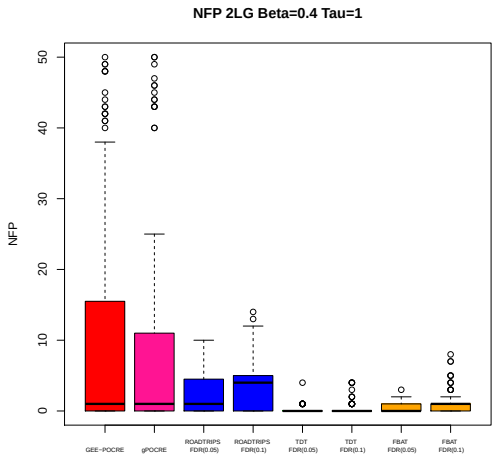
(a) $\tau = 0.9$: True Positives



(b) $\tau = 0.9$: False Positives

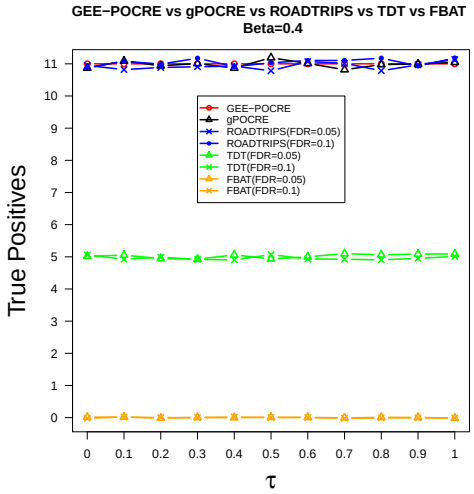


(c) $\tau = 1$: True Positives

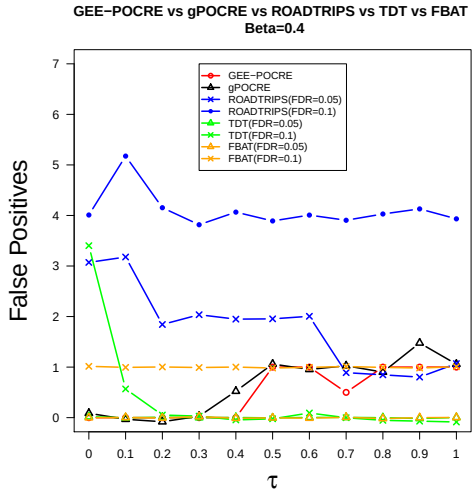


(d) $\tau = 1$: False Positives

Figure 3.1.: Boxplot comparisons among GEE-POCRE, gPOCRE, ROADTRIPS, TDT and FBAT in the simulation study of eleven SNPs in two linkage groups



(a) Median true positives in the simulation study of eleven SNPs in two linkage groups. Effect size $\beta = 0.4$.



(b) Median false positives in the simulation study of eleven SNPs in two linkage groups. Effect size $\beta = 0.4$

Figure 3.2.: Median true positives and false positives plots in the simulation study of eleven SNPs in two linkage groups.

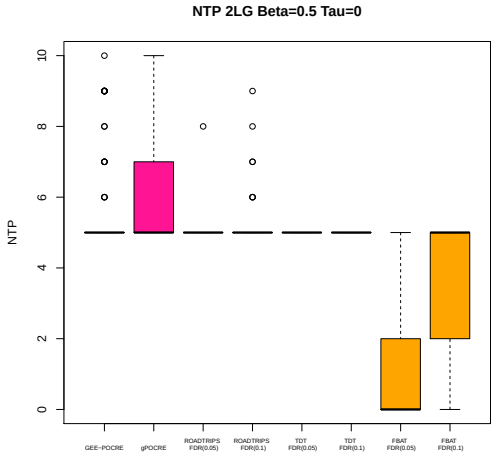
Case II. Fifteen Causal SNPs Existing in Two Linkage Groups

Denote the 15 causal SNPs as $X_1, \cdots, X_{15}$. We then simulate phenotypic values from the following logistic model, $\text{logit}(E[Y]) = \mu + 0.5\sum_{j=1}^{28} X_j$, where $\text{var}(Y) = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$, A is diagonal matrix with the variances of Y as the diagonal elements, and $\mathbf{R}$ is a correlation matrix defined as $\mathbf{R} = (1-\tau)\mathbf{I} + \tau\mathbf{K}$. We consider different values of τ, ranging from 0 to 1 with increment 0.1. For each τ, we simulate 100 data sets. Two sets of results are reported with ROADTRIPS, TDT and FBAT by controlling FDR at 0.05 and 0.1 respectively.

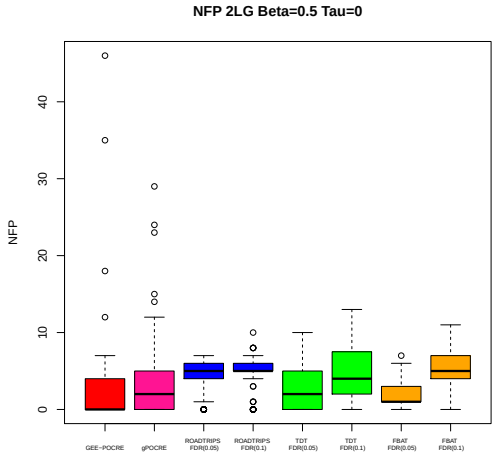
The boxplots of true positives and false positives in the case II simulation study among all the compared methods are shown in Figure 3.3. GEE-POCRE, gPOCRE, ROADTRIPS, TDT are able to find five true SNPs compared with FBAT which has less true detections and more variations. A greatly decreased false positives is observed in GEE-POCRE and its performance is relatively stable. We also observe a more obvious tendency that GEE-POCRE has less variation in terms of both true positives and false positives across different τ when compared with gPOCRE. Figure 3.4 summarizes the median true positives and false positives of all compared methods across different τ. It clearly shows that GEE-POCRE performs the best in all our simulation studies.

3.4 The NIA-Late Onset Alzheimer's Disease Application

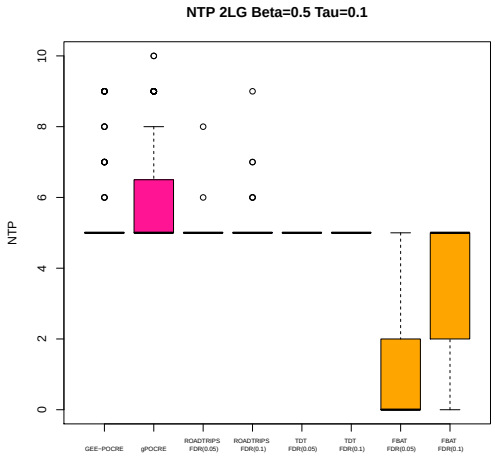
Alzheimer disease (AD) is the most common neurodegenerative disorder in the elder population that affects millions of Americans. In 2002, The National Institute of Aging (NIA) started a NIA-LOAD study, which contains families with two or more siblings with the late onset form of alzheimer's disease and a cohort of unrelated controls similar in age and ethnic background (Lee *et al.* 2008). Among all the samples, a neuropathological criteria is used to diagnose AD and controls are defined as individuals without noticing any loss of memory and pass the neuropsychological test. We clean the data by removing samples with missing rate more than 10%, SNPs



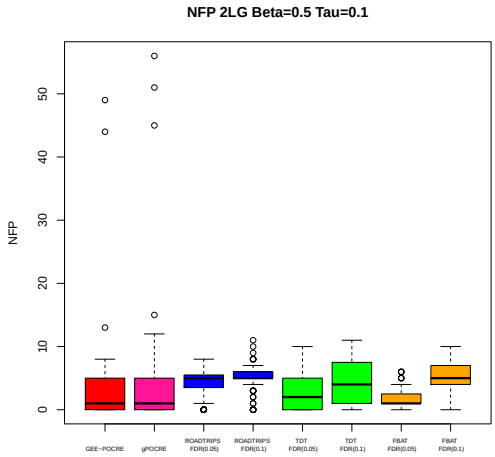
(a) $\tau = 0$: True Positives



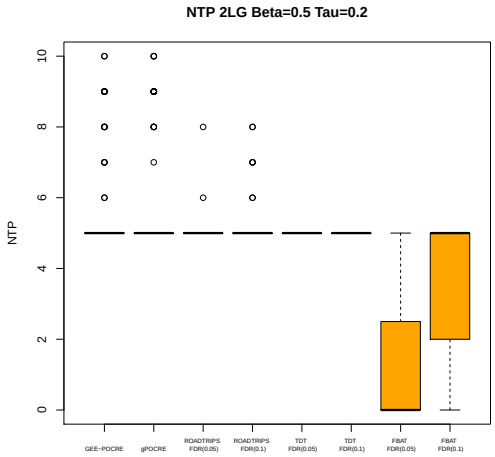
(b) $\tau = 0$: False Positives



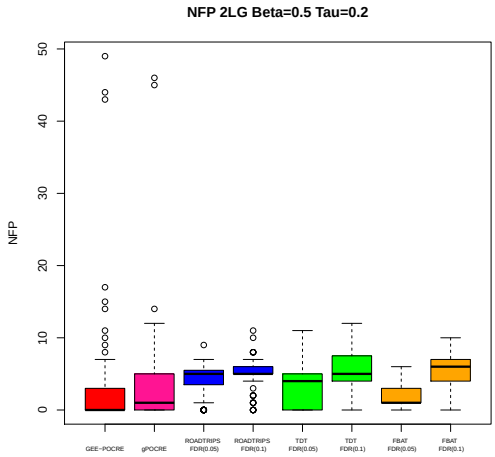
(c) $\tau = 0.1$: True Positives



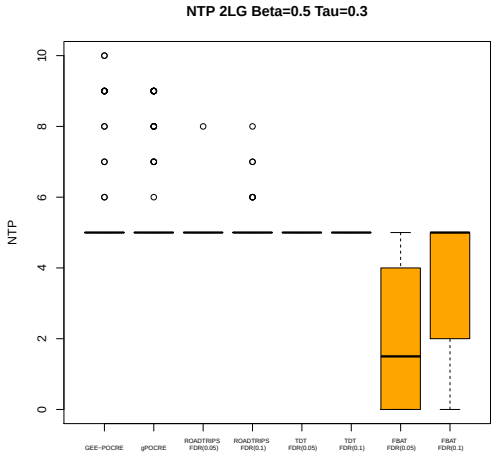
(d) $\tau = 0.1$: False Positives



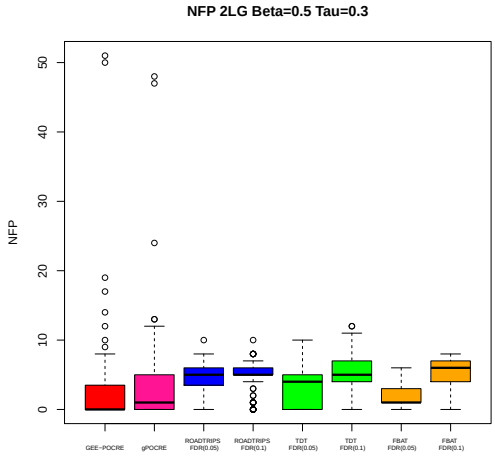
(e) $\tau = 0.2$: True Positives



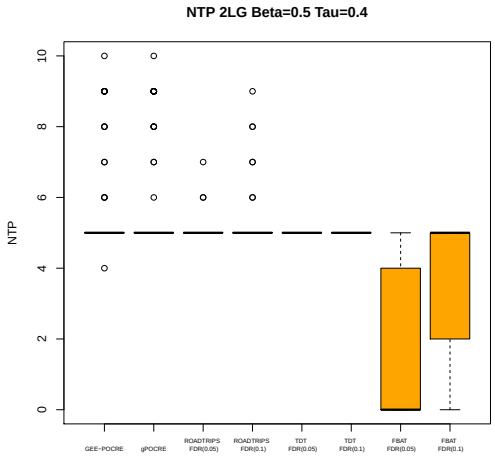
(f) $\tau = 0.2$: False Positives



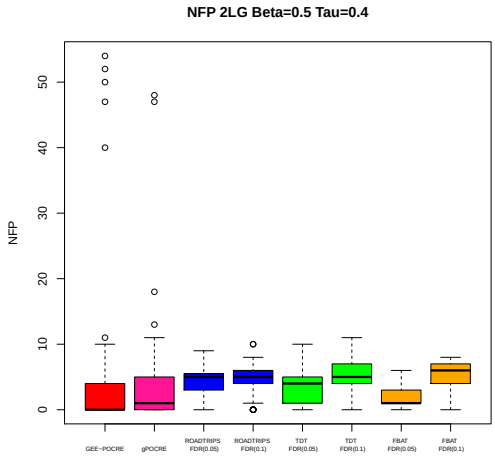
(a) $\tau = 0.3$: True Positives



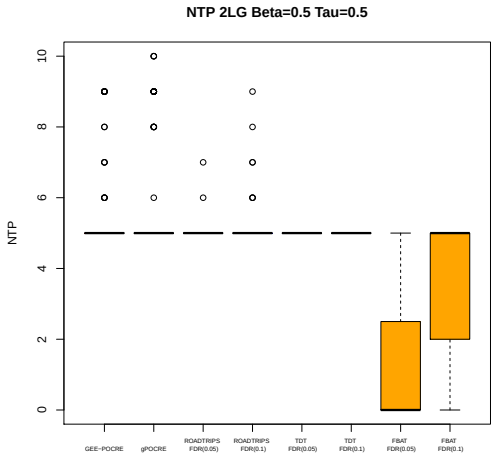
(b) $\tau = 0.3$: False Positives



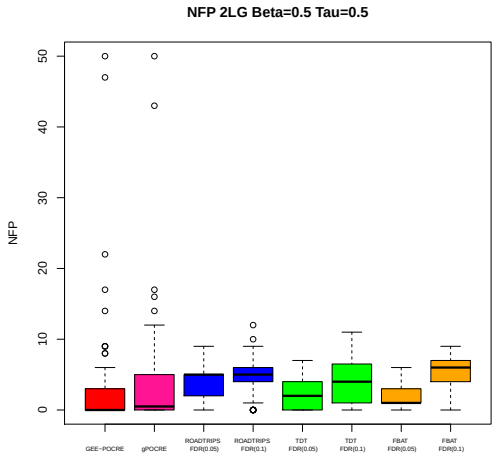
(c) $\tau = 0.4$: True Positives



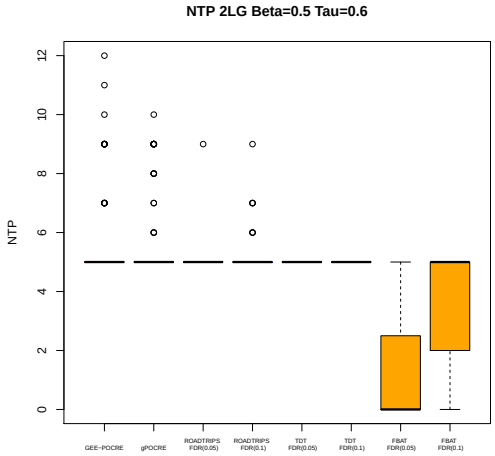
(d) $\tau = 0.4$: False Positives



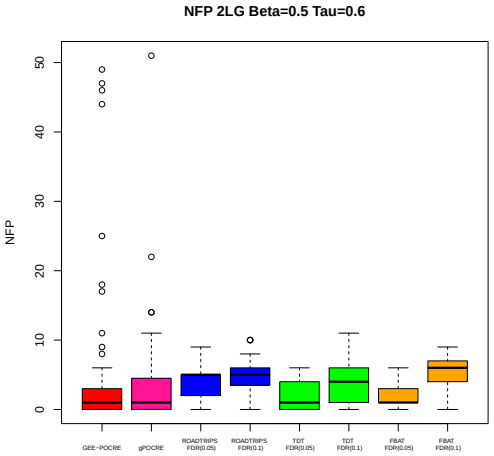
(e) $\tau = 0.5$: True Positives



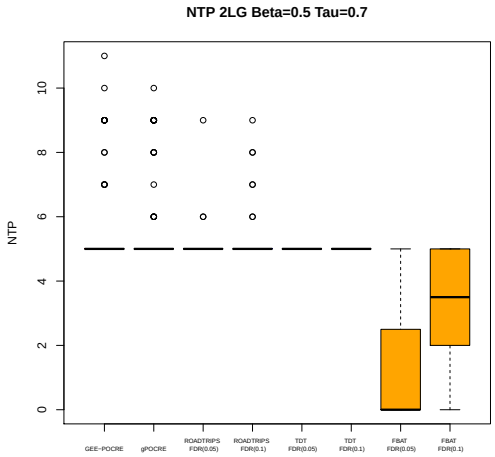
(f) $\tau = 0.5$: False Positives



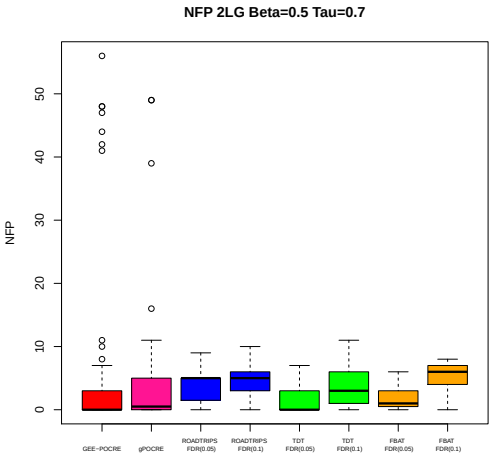
(a) $\tau = 0.6$: True Positives



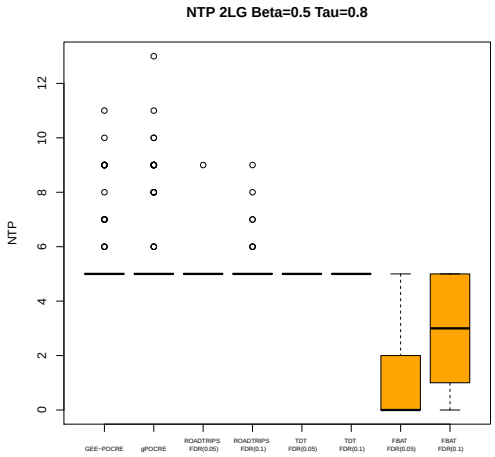
(b) $\tau = 0.6$: False Positives



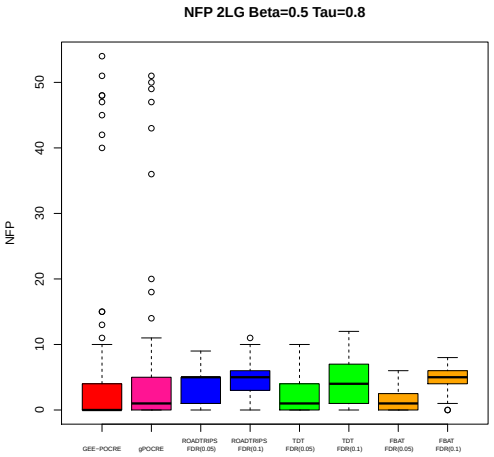
(c) $\tau = 0.7$: True Positives



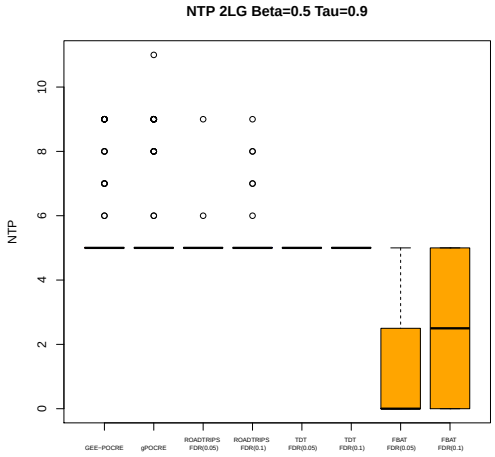
(d) $\tau = 0.7$: False Positives



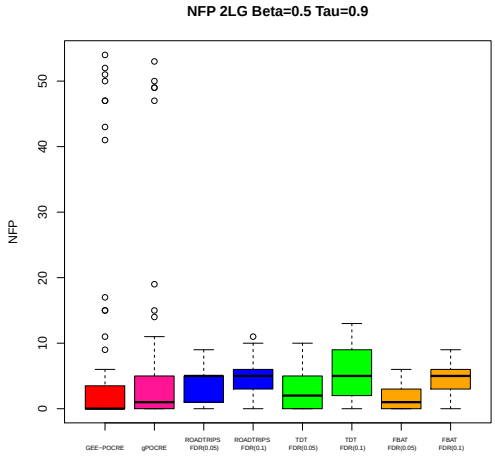
(e) $\tau = 0.8$: True Positives



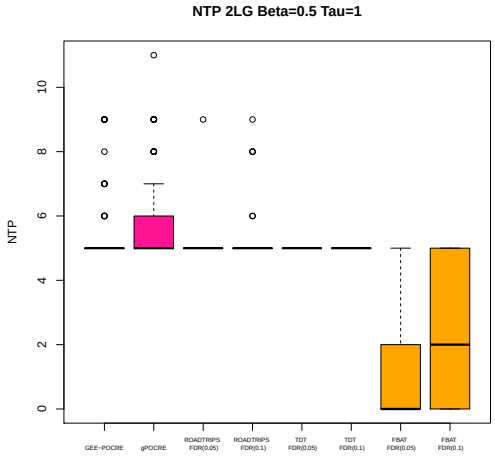
(f) $\tau = 0.8$: False Positives



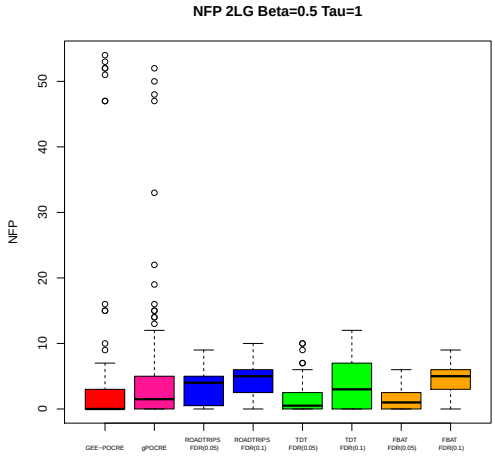
(a) $\tau = 0.9$: True Positives



(b) $\tau = 0.9$: False Positives

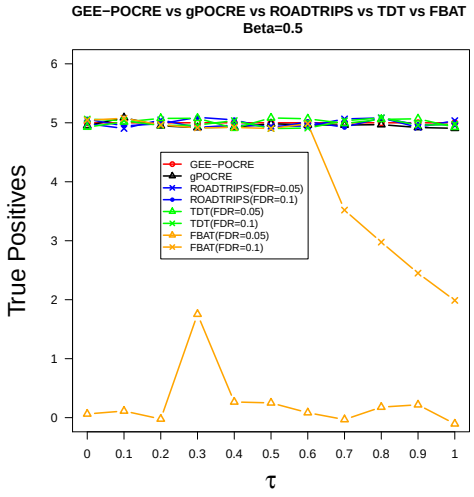


(c) $\tau = 1$: True Positives

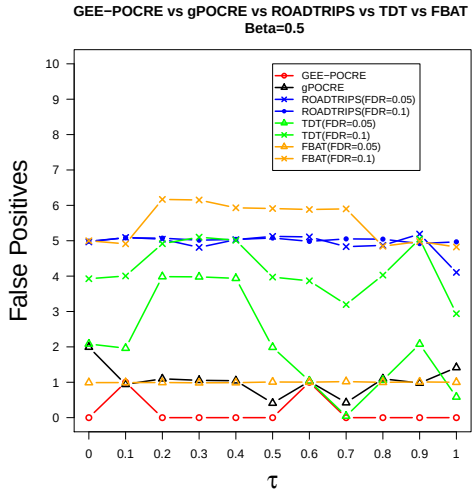


(d) $\tau = 1$: False Positives

Figure 3.3.: Boxplot comparisons among GEE-POCRE, gPOCRE, ROADTRIPS, TDT and FBAT in the simulation study of fifteen SNPs in two linkage groups



(a) Median true positives in the simulation study of fifteen SNPs in two linkage groups. $\beta = 0.5$



(b) Median false positives in the simulation study of fifteen SNPs in two linkage groups. $\beta = 0.5$

Figure 3.4.: Median true positives and false positives plots in the simulation study of fifteen SNPs in two linkage groups.

with missing rate per sample more than 10% and minor allele frequency less than 0.05. Overall, a total of 2545 samples that are genotyed at 532,795 SNPs are left for further analysis. We then apply GEE-POCRE to this data set using 10-folds cross validation.

Table 3.1.: GEE-POCRE results on late onset Alzheimer's disease

Chromosome	SNP	Physical Position	Beta	P Value*	Gene
5	rs1477280	160,691,741	6.49E-02	6.65E-03	*ATP10B*
17	rs4789374	76,916,278	-2.96E-02	3.95E-04	*MGAT5B*
18	rs3888795	11,863,900	1.27E-01	2.84E-02	*GNAL*
19	rs2075650	4,539,619	1.98E-01	1.63E-31	*APOE*

*p-values are calculated using the 10-split method

The results are shown in Table 3.1. In hypothesis testing, p-value is a popular indicator to quantify statistical significance. We then try to assign a p-value to the SNPs we find. Recently, Meinshausen et al. (2009) has proposed a multi-split method to assign statistical significance and construct p-values for high-dimensional analyses where the number of predictors may be much larger than the sample size. In each split, the data is divided into two parts, GEE-POCRE uses the first part and builds a statistical model, then a classical variable selection technique is applied to the selected variables using the data from the second part. The method has the property of asymptotic error control and model selection consistency. Here we apply multi-split method with a total of 10 splitting. Reported are the ones with a p-values less than 0.05.

There are four SNPs identified by GEE-POCRE. The one locates on chromosome 19 that is associated with *APOE* shows compelling evidence of association with LOAD and has been confirmed in other studies (Liu et al. 2013). *GNAL* locus also known as *DYT25* is recorded in NCBI gene database as a functional gene that encodes a stimulatory G protein alpha subunit and is wildly expressed in the central nervous

system. This gene is found to be susceptible to schizophrenia, now we have good evidence to suspect it may also influence the late onset alzheimer's disease. The SNPs rs1477280 and rs4789374 that resides in the gene region *ATP10B* and *MGAT5B* are also statistically significant, their impacts on alzheimer's disease is not clear yet, and needs further investigation, e.g. looking into the GO of these two genes, and checking whether they have been reported as eQTLs in any other related studies.

3.5 Conclusion

In this chapter, we propose GEE-POCRE by extending penalized orthogonal components regression in the generalized estimating equations model. Incorporating the kinship matrix into modeling the variance covariance structure in GEE as well as applying penalization functions when constructing orthogonal components, GEE-POCRE effectively handles the family structure when simultaneously does the variable selection and estimation in high-dimensional data analysis. Simulation studies and a real data analysis are carried out to evaluate and compare the performance of our proposed novel approach and some other popular existing methods including TDT, FBAT, ROADTRIPS and gPOCRE. Both simulations and the real data example demonstrate a good performance of GEE-POCRE. Particularly, GEE-POCRE has the same or more power than gPOCRE, ROADTRIPS, TDT, and FBAT, and much lower false positives.

4. SUMMARY AND FUTURE WORK

4.1 Summary

In this dissertation, we propose fPOCRE which extends the penalized orthogonal components regression in the linear mixed model with continuous responses, and the GEE-POCRE that constructs orthogonal components through a penalization function assuming the generalized estimating equations model for binary traits. Both the algorithms take the kinship matrix into the modeling process and are multi-locus methods compared to popular existing methods that are essentially single-locus analyses.

Utilizing the linear mixed model, fPOCRE iteratively estimates the ratio between the genetic and the residual variation τ and the fixed SNP effects β separately. First assuming a known τ, the SNP effects are estimated with sparsity, then the τ is updated with the estimated fixed SNP effects through the profiled log-likelihood. The algorithm iterates between updating τ and estimating SNP effects until τ converges. fPOCRE can work on really large data sets due to its computational efficiency. Two simulation studies show the superior performance of fPOCRE in terms of high power and low false positives, while EMMAX always select a large number of variables with many that are false, and MLMM are more conservative with a low detecting rate, leading to loss of power. Finally, the results in real data applications are concordant with the results in our simulation studies.

Assuming the generalized estimating equations model, GEE-POCRE is developed by incorporating the kinship matrix into the variance covariance structure, taking turns to construct orthogonal components and estimate the unknown parameter τ in the working correlation matrix. In our simulation studies, GEE-POCRE outperforms the popular family-based algorithms, TDT, FBAT, ROADTRIPS that test one SNP at a time, as well as a multi-locus method gPOCRE that assumes unrelated samples.

A LOAD data analysis further confirms the conclusion we draw from the simulation studies.

4.2 Future Works

4.2.1 Extension of Model Selection Criteria in GEE-POCRE

Model selection is an important issue in any type of data analysis. Currently, GEE-POCRE uses 10-folds cross-validation to select the optimal tuning parameter λ. However, it is computationally expensive to perform cross-validation in large data sets, moreover, some observations may never be in the training data set and this may bias the resultant model. Other powerful and easily implemented model selection techniques are worth to explore, such as AIC, BIC and eBIC. However, GEE-POCRE is not a likelihood based model. Directly applying the information criteria would not be appropriate since there is no likelihood defined. Pan (2001) proposed the QIC, a quasi-likelihood information criteria that modifies the the well-known Akaike Information Criteria, where the likelihood was replaced by the quasi-likelihood and the penalty term is also adjusted accordingly. Investigating on the usage of QIC in our proposed GEE-POCRE would be beneficial.

4.2.2 Extension to Other GLMs

Even though the task of this dissertation is variable selection in GWAS with family structure for both normally distributed responses and binary responses, there are many other possible regression models to which we could consider to extend our proposed algorithm. We list a few possible options below.

First, assuming researchers want to perform a GWA study on the number of strokes occurring to a patient within one year period. The phenotype of interest is count data, therefore the log-linear model which assumes the responses follow Poisson distribution is suitable for such type of analysis given the samples are unrelated.

However, if the genetic relatedness among study subjects is a concern, our proposed GEE-POCRE algorithm with log-link function would be an alternative solution and is worth to investigate. Second, clinical data over the years are commonly collected in GWAS. Knowing that GEE was initially proposed for longitudinal data analysis, it is natural to consider extending GEE-POCRE in the context of longitudinal study.

4.2.3 Extension to Multiple Traits

Genome-wide association studies are usually measured with multiple traits, among them, many are highly correlated. For example, considering a study on the bone mineral density (BMD), many sub-traits are measured to evaluate BMD. Analyses using these sub-traits marginally inevitably loses some essential information among these multiple correlated traits. An integrative method that borrows strength across traits, as well as assumes a multi-locus model that considers the joint effects of multiple genetic variants would be promising. Even though the example we discuss above is related to the biological study, however, the problem of high-dimensional data analysis with structured samples are everywhere, including social behavior studies, psychology and sociology where multiple correlated measurements are frequently observed.

REFERENCES

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* Wiley-Interscience.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control 19*, 716-723.

Albert, A. (1984). On the existence of maximum likelihood estimatees in logistic regression models. *Biometrika 71*, 1-10.

Astle, W., and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science 24*, 451-471.

Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *The Annals of Statistics 32*, 870-897.

Baxter, I., Brazelton, J. N., Yu, D., Huang, Y. S., Lahner, B., Yakubova, E., Li, Y., Bergelson, J., Borevitz, J. O., Nordborg, M., Vitek, O. and Salt, D. E. (2010). Coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter *AtHKT1;1. PLoS Genetics, 16(11)*: e1001193 doi:10.1371/journal.pgen.1001193.

Benyamin, B., Visscher, P. M. and McRae, A. F. (2009). Family-based genome-wide association studies. *Pharmacogenomics, 10*, 181-190.

Boulesteix, A.-L., and Strimmer, K. (2006). Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data. *Briefings in Bioinformatics, 8*, 32-44.

Breiman, L. (1995). Better subset regression using the nonnegative garotte. *Technometrics 37*, 373-384.

Breiman, L.(1996). Heuristics of instability and stabilization in model selection. *The Annal of Statistics, 24*, 2350-2383.

Chen, J. and Chen, Z.(2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika, 95*, 759-771.

Chen, M., Liu, X., Wei, F., Larson, M. G., Fox, C. S., Vasan, R. S., and Yang, Q. (2011). A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genetic Epidemiology, 35*, 650-657.

Chung, H., and Keles, S. (2010). Simultaneous dimension reduction and variable selection with Sparse Partial Least Squares. *Journal of Royal Statistical Society - Series B (Statistical Methodology), 72*, 3-25.

Chung, D., and Keles, S. (2010). Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology, 9*, Issue 1, Article 17.

Cupples, L.A., Arruda, H.T., Benjamin, E.J., *et al.* (2007). The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics 8*, (Suppl1):S1.

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

Donoho, D.L. and Johnstone I.M. (1994). Spatial adaptation via wavelet shrinkage. *Biometrika 81*, 425-455.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science 18*, 71-103.

Dunlop, D. D. (1994). Regression for longitudinal Data: A bridge from Least Sqaure Regression. *The American Statistician 48*, 299-303.

Efron, B., Hastie, T., Johnstone, I. M. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics 32*, 407-499.

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

Fan, J.and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B 70*, 849-911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research 10*, 2013-2038.

Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*, 101-148.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27-38.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1-22.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 2409-2419.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., Vilhjalmsson, B. J., Nordborg, M. Borevitz, J. O. and Bergelson, J. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics, 44*, 212-216.

Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics, 2*, 211-228.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystack: empirical Bayes estimates of possibly sparse sequence. *The Annals of Statistics, 32*, 1594-1649.

Johnstone, I. M. and Silverman, B. W. (2005). Ebayes Thresh: R programs for empirical Bayes thresholding. *Journal of Statistical Software, 12*, 1-38.

Jorgensen, M.A. (1994). Tail functions and iterative weights in binary regression. *The American Statistician 48*, 230-234.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C. and Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics, 178*, 1709-1723.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics, 42*, 348-354.

Knowler, W. C., Williams, R. C., Pettitt, D. J., and Steinberg, A. G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: An association in American Indians with gentic admixture. *Am. J. Hum. Genet. 43*, 520-526.

Laird, NM., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family based tests of association. *Genetic Epistemology. 19*, S36-S42.

Lane, S. (2007). Generalized Estimating Equations for Pedigree Analysis *Thesis*

Lee, J. H. *et al.* (2008). Analyses of the national institute on aging Late-Onset Alzheimer's disease family study: implication of additional loci. *Archives of Neurology, 65*, 1518-1526.

Legarra, A., Granie, C. R., Manfredi, E. and Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics, 180*, 611-618.

Levy, M. and Wang, V. (2013). The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective, *Lancet, 27*, 61752-61753.

Liang, G. and Yu, B. (2003). Maximum pseudo likelihood estimation in network tomography. *IEEE Transactions on Signal Processing 51*, 2043-2053.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 78*, 13-22.

Lin, Y., Zhang, M., and Zhang, D. (2014). Generalized orthogonal components regression for high dimensional genearlized linear models. *Purdue University Technical Report*

Liu, C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisums and therapy. *Nature Reviews Neurology, 9*, 106-118.

Marx, B. (1996). Iteratively reweighted partial least squares estimation for generalized linear model regression. *Technometrics, 38*, 374-381.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncerntainty and challenges. *Nature Review Genetics, 9*, 356-369.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall.

Meier, L., van de Geer, S. and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B 70(1)*, 53-71.

Meinshausen, N., Meier, L. and Buhlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*, 1671-1681.

Meng, X. (2008). Discussion: One-step sparse estimates in non-concave penalized likelihood models: Who cares if it is a white cat or a black cat?. *Annals of Statistics 36*, 1542-1552.

Newton, M.A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*, 155-176.

Ott, J., Kamatani, Y. and Lathrop, M. (2011). Family-based designs for genome-wide association studies *Nature reviews 12*, 465-474.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations *Biometrics 57*, 120-125.

Pinheiro, J. C. and Bates, D. M. (2000). *Am J Hum Genetics 67*, 170-181.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principle components analysis corrects for stratificaiton in genome-wide association studies. *Nature Genetics 38*, 904-909.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New apporaches to population stratification in genome-wide association studies. *Nature Review Genetics 11*, 459-463.

Pritchard, J. K., Stephens, M., Rosenberg, N., and Donnelly, P. (2000). *Mixed Effects Models in S and S-Plus.*New York: Springer-Verlag.

Rakovski, D., and Laird, NM. (2000). A unified approach to adjusting association tests for population admixture with arbitary pedigree structure and arbitary mssing marker information. *Human Heredity 504*, 227-233.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461-464.

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q. and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics, 44*, 825-830.

She, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics 3*, 384-415.

Spielman, R. S., Mcginnis, R. E., and Evens, W. J. (1993). Transmission test for linnkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). *Ammerican Journal of Human Genetics 52*, 506-516.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B 58*, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Socity Series B 67*, 91-108.

Tiwari H. K., Sloan, J. B., Wineinger, N., Padilla, M. A., Vaughan, L. K., and Allison, D. B. (2008). Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Human Heredity 66*, 67-86.

Valdar, W., Solberg, L. C., Gauguier, D. Burnett, S., Klenerman, P., Cookson, W. O., Taylor, M. S. Rawlins, J. N. P., Mott, R., and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics, 38*, 879-887.

Varin, C., Reid, N. and Firth, F. (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5-42.

Vinzi, V. E., Chin, W. W., Henseler, J., and Wang, H. (2010). *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Berlin: Springer.

Wang, X., Park, T. and Carriere, K.C. (2010). Variable selection via combined penalization for high-dimensional data analysis. *Computational Statistics and Data Analysis 54*, 2230-2243.

Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika 61*, 439-447.

Wold, H. (1975). Soft modeling by latent variables: the nonlinear iterative partial least squares approach. *Perspectives in probability and statistics, papers in honor of M. S. Bartlett. London: Academic Press*, 117-142.

Wold, S., Ruhe, A., Wold, H., Dunn, W.J. (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing, 5*, S97.

Wu, X., Kan, D., Cooper, R., and Zhu, X. (2005). Identifying genetic variation affecting a complex trait in simulated data: a comparision of meta-analysis with pooled data analysis. *BMC Genetics, 6*, 735-743.

Yamada, K., Iwayama, Y., Hattori, E., Iwamoto, K., Toyota, T., Ohnishi, T., Ohba, H., Maekawa, M., Kato, T., and Yoshikawa, T. (2011). Genome-wide association study of schizophrenia in Japanese population. *PLoS One, 6*, doi: 10.1371.

Yu, J., Pressoir. G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F. McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. and Buckler, E. S. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics, 38*, 203-208.

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

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continous outcomes. *Biometrics 42*, 121-130.

Zeger, L.P. and Rentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika 77*, 642-648.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*, 894-942.

Zhang, D., Lin, Y. and Zhang, M. (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics 3*, 781-796.

Zhang, M., Zhang, D. and Wells, M. T. (2010). Generalized thresholding estimators for high-dimensional location parameters. *Statistica Sinica 20*, 911-926.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*, 894-942.

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics 42*, 355-360.

Zou, H. and Hastie, H. (2005). Regularization and variable selection via the elastic net. *Journal of Computational and Graphical Statistics 15*, 265-286.

Zou, H. (2006). The adaptive lasso and its oracle propertie. *Journal of the American Statistical Association 101*, 1418-1429.

VITA

VITA

Libo Wang was born in China. She received a B.S. in Applied Mathematics from Dalian University of Technology, China in 2006. Later, She studied Computational Finance, from 2006 to 2008 at department of mathematics of Purdue University. In 2008, she entered department of Statistics under the supervision of Prof. Dabao Zhang and Min Zhang, received the master of statistics in 2010. Her academic interests include linear mixed model, statistical modeling and model selection (especially in massive data) and computational statistics.

Libo Wang's publications for her research work at Purdue University include:

Publications(Published)

- Wang, L., Pungpapong, V., Lin, Y., Zhang, M., and Zhang, D. (2011). Genome-wide case-control study in GAW17 using coalesced rare variants. BMC Proceeding, 5 (Suppl 9), S110.
- Wang, L., Athinarayanan, S., Chalasani, N., Zhang, M., and Liu, W. (2014) Fatty Acid Desaturase 1 (FADS1) Gene Polymorphisms Control Human Hepatic Lipid Composition. Hepatology, doi:10.1002/hep.27373.
- Replogle. R. A., Li, Q., Wang, L., Zhang, M., and Fleet, J. C. (2014) Gene-by-Diet Interactions Influence Calcium Absorption and Bone Density in Mice. Journal of Bone and Mineral Research, 29(3): 657-665.
- Gamazon E.R., Innocenti F., Wei R., Wang L., Zhang M., Mirkov S., Ramirez R., Huang R.S., Cox N.J., Ratain M.J., and Liu W. (2013). A genome-wide integrative study of microRNAs in human liver. BMC Genomics, 14: 395.
- Thakur, K. K., Pant, G.R., Wang, L., Hill, C.A., Pogranichniy, R.M., Mannadhar, S., and Johnson, A. J. (2012). Seroprevalence of Japanese Encephalitis

Virus and Risk Factors Associated with Seropositivity in Pigs in Four Mountain Districts in Nepal. Zooneoses and Public Health, 59(6): 393-400

- Bailey-Wilson, J.E., Brennan, J.S., Bull, S.B., Culverhouse, R., Kim, Y., Jiang, Y., Jung, J., Li, Q., Lamina, C., Liu, Y., Mgi, R., Niu, Y.S., Simpson, C.L., Wang, L., Yilmaz, Y.E., Zhang, H., and Zhang, Z.(2011) Regression and data mining methods for analyses of multiple rare variants in the Genetic Analysis Workshop 17 mini-exome data. Genet Epidemiol. 35 (Suppl 1)
- Pungpapong, V., Wang, L., Lin, Y., Zhang, D., and Zhang, M. (2011). Genome-wide association analysis of GAW17 data using an empirical Bayes variable selection. BMC Proceeding, 5 (Suppl 9), S4.
- Lin, Y., Zhang, M., Wang, L., Pungpapong, V., Fleet, J.C., and Zhang, D. (2009). Simultaneous genome-wide association studies of anti-CCP in rheumatoid arthritis using penalized orthogonal-components regression. BMC Proceedings, 3 (Suppl 7), S20.
- Zhang, M., Lin, Y., Wang, L., Pungpapong, V., Fleet, J.C., and Zhang, D. (2009). Case-control genome-wide association study of rheumatoid arthritis from GAW16 using POCRE-LDA. BMC Proceedings, 3 (Suppl 7), S17.

Publications(Submitted & In Preparation)

- Replogle. R. A., Wang, L., Zhang, M., and Fleet, J. C. (2014) Identification of novel genetic loci affecting serum vitamin D metabolite levels and the response of serum 1,25 dihydroxyvitamin D levels to dietary Ca restriction in the BXD recombinant inbred mouse panel. Journal of Bone and Mineral Research, submitted.
- Wang, L., Zhang, M., and Zhang, D. (2014) Identification of genomic factors using family-based association studies, in preparation.
- Wang, L., Zhang, M., and Zhang, D. (2014) An efficient method for case-control genome-wide associations with family structure, in preparation.

- Zhang, M., Zhang, D., Wang, Q., Wang, L. and Guan, L. (2014) A GWAS Follow-up Study in Han Chinese Population: Common Variants on 17q25 Conferring Risk of Schizophrenia, Molecular Psychiatry, in preparation.