Statistical Methods for Analyzing Large-Scale Biological Data

by

Rounak Dey

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Associate Professor Seunggeun Lee, Chair
Professor Gonçalo Abecasis
Associate Professor Hyun Min Kang
Associate Professor Cristen Willer

Rounak Dey

deyrnk@umich.edu

ORCID iD: 0000-0002-6540-8280

This is for you, Lal

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor Shawn Lee for his support and guidance throughout my doctoral studies. He always motivated me whenever I had any doubt in my research. His insights and problem solving abilities have always inspired me, and helped me pave my path as a researcher. I would also like to thank other members of the dissertation committee, Gonçalo Abecasis, Hyun Min Kang, and Cristen Willer for their help and guidance in different collaborative projects, and for their assistance in completing this thesis. I am also grateful to Michael Elliott for his support, and the opportunity to work with him during the first two years of my PhD.

Here, I would like to thank my collaborators for their helpful contributions: Kwangsik Nho for his help with the Alzheimer's Disease Neuroimaging Initiative data in my research on partial least squares, Ellen Schmidt and Gonçalo Abecasis for maintaining and providing leadership regarding the use of the Michigan Genomics Initiative data, and Jonas Nielsen, Lars Fritsche and Cristen Willer for preparing the phenome for the UK Biobank data. Thanks also go to Wei Zhou and Huanhuan Zhu for their helpful comments and suggestions to improve my research on biobank-based meta-analysis methods, and to Hyun Min Kang for his help on implementing my methods on the EPACTS software. All my research works were supported by NIH Grants R00HL113164 and R01HG008773. My access to the UK Biobank data was provided under the application number 24460.

I am thankful to my friends in high school and in ISI: Rejaul Karim, Kushal

Kumar Dey, Moumanti Podder, La Krusade, Satyajit Ghosh, Rahul Rahaman, Riddhiman Bhattacharya, Sabyasachi Bera, Abhirup Mondal, Sayar Karmakar, Arkajyoti Bhattacharya, Arka Bhattacharya, Chinmoy Bhattacharjee, Deepan Basu, Rudradev Sengupta, Tamal Kumar De, Shrijita Bhattacharya, Avijit Kumar Dutta, Abhishek Kumar, Sayak Chowdhury, Soudeep Deb, Angshuman Roy, Narayan Bose, Biswarup Bhattacharya, and many others, for sharing their lives with me and for being there for me whenever I needed them. Thanks go to my ISI seniors and juniors in Michigan, especially Sayantan Das, Aritra Guha, Diptavo Dutta, Anwesha Bhattacharya, Debarghya Mukherjee, Moulinath Banerjee and Bhramar Mukherjee for the dinner parties, late night hangouts, and so many lifelong cherish-able memories made here in Ann Arbor. I would also like to thank my friends and colleagues in the Biostatistics department: Jingchunzi Shi, Alan Kwong, Wei Zhou, Yumeng Li, Sheng Qiu, Paul Imbriano, Jessica Lehrich, Tingting Zhou, Sai Dharmarajan, Tian Gu, Pranav Yajnik and others, for their helpful suggestions on my research and the fun hangouts we had.

Finally, I would like to express my love and gratitude to my parents, cousins, grandmothers, uncles and aunts, for their love and support throughout my life, and to Lal, my friend, philosopher, and guide, for teaching me the expansion of $(a+b)^2$.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

LIST OF ABBREVIATIONS

AD Alzheimer's disease

ADNI Alzheimer's Disease Neuroimaging Initiative

AF allele frequency

AR auto-regressive

CEU Utah residents with Northern and Western European ancestry

CGF cumulant generating function

CPU central processing unit

CV cross-validation

EF executive functioning scores

EHR electronic health records

EMCI early mild cognitive impairment

ESD empirical spectral distribution

GC genotype count-based method

GSP generalized spiked population

GWAS genome-wide association study

HRC Haplotype Reference Consortium

ICD International Classification of Disease

LMCI late mild cognitive impairment

LSD limiting spectral distribution

MAC minor allele count

MAF minor allele frequency

MEM memory scores

MGI Michigan Genomics Initiative

MHC Major Histocompatibility Complex

MLR multiple linear regression

MRI magnetic resonance imaging

MSE mean-squared error

MSEP mean-squared error of prediction

MTL medial temporal lobe

NIH National Institutes of Health

OLS ordinary least squares

PC principal component

PCA principal component analysis

PheWAS phenome-wide association study

PLS partial least squares

QQ quantile-quantile

RHS right-hand side

SNP single nucleotide polymorphism

SP spiked population

SPA saddlepoint approximation

SPLS sparse PLS

TPLS two-stage PLS

TSI Toscans in Italy

UHD ultra high-dimensional

ABSTRACT

With the development of high-throughput biomedical technologies in recent years, the size of a typical biological dataset is increasing at a fast pace, especially in the genomics, proteomics and metabolomics literatures. Typically, these large datasets contain a huge amount of information on each subject, where the number of subjects can range from small to often extremely large. The challenges of analyzing these large datasets are twofold, namely the problem of high-dimensionality, and the heavy computational burden associated with analyzing them. The goal of this dissertation is to develop statistical and computational methods to address some of these challenges in order to provide researchers with analytical tools that are scalable to handle these large datasets, as well as able to solve the issues arising from high-dimensionality.

In Chapter II, we study the asymptotic behaviors of principal component analysis (PCA) in high-dimensional data under the generalized spiked population model. We propose a series of methods for the consistent estimation of the population eigenvalues, angles between the sample and population eigenvectors, correlation coefficients between the sample and population principal component (PC) scores, and the shrinkage-bias adjustment for the predicted PC scores.

In Chapter III, we investigate the over-fitting problem of partial least squares (PLS) regression with high-dimensional predictors, which can result in the predicted and observed outcomes being almost identical, even when the outcome is independent of the predictor. We further discuss a shrinkage-bias problem similar to the shrinkage-bias in high-dimensional PCA, and propose a two-stage PLS (TPLS) method that can address both of these problems.

In Chapter IV, we focus on the large-scale genome-wide or phenome-wide association studies (GWASs or PheWASs) of the electronic health records (EHR) or biobank-based binary phenotypes. Due to the severe case-control imbalance in most of the EHR or biobank-based binary phenotypes, the existing methods cannot provide a scalable and accurate way to analyze them. We develop a computationally efficient single-variant test, that is ~ 100 times faster than the state of the art Firth's test, and can provide well-calibrated p values even for phenotypes with extremely unbalanced case-control ratios. Further, our test can adjust for non-genetic covariates, and can retain similar power as the Firth's test.

In Chapter V, we show that due to the severe case-control imbalance in most of the biobank-based binary phenotypes, applying the traditional Z-score-based method to meta-analyze the association results across multiple biobank-based association studies, can result in conservative or anti-conservative p values. We propose two alternative meta-analysis methods that can provide well-calibrated meta-analysis p values, even when the individual studies are extremely unbalanced in their case-control ratios. Our first method involves sharing an approximation of the distribution of the score test statistic from each study using cubic Hermite splines, and the second method involves sharing the overall genotype counts from each study.

In summary, the purpose of this dissertation is to develop statistical and computational methods that can efficiently utilize the ever-growing nature of modern biological datasets, and facilitate researchers by addressing some of the problems associated with the high-dimensionality of the datasets, as well as by reducing the heavy computational burden of analyzing these large datasets.

CHAPTER I

Introduction

In recent years, the size of a typical biological dataset has been increasing at a very fast pace, thanks to the drastic developments in low-cost high-throughput biomedical technologies. In the field of genomics, the current genotyping and imputation technologies (*Marchini and Howie*, 2010; *Das et al.*, 2016) allow for genotyping tens of millions of variants at a very low cost. Major developments in high-throughput drug discovery have led to the generation of a vast amount of information at a very low cost on the transcriptome, proteome, glycome and metabolome (*Howbrook et al.*, 2003; *Sun et al.*, 2013). Modern functional and anatomical neuroimaging techniques (*Monchi et al.*, 2008; *Williams and Henson*, 2018) have allowed efficient and low-cost imaging of the entire brain in great detail. Recent studies have also been focusing on generating these large datasets from different sources, and modeling them together. For example, the availability of genotype data, along with electronic health records (EHR)-based phenotypes in biobanks (*Bycroft et al.*, 2017; *Krokstad et al.*, 2013), have enabled us to perform genome-wide association studies (GWASs) in phenome-wide scales (*Hebbring*, 2014; *Verma et al.*, 2018; *Fritsche et al.*, 2018). In the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, the collection of genotyping and whole genome sequencing data along with the neuroimaging data, has enabled us to gather detailed insight on the influence of the genetic factors on the changes in the human brain at

the structural, functional and molecular levels, that can eventually lead to the onset of the Alzheimer's disease (*Saykin et al.*, 2010; *Moon et al.*, 2015; *Shen et al.*, 2014).

Even though it is becoming easier to gain enormous amount of information at a relatively lower cost, these large datasets are posing new kinds of challenges to the research community. Primarily, the challenges are twofold. Firstly, analyzing such datasets requires addressing the problem of high-dimensionality, as a lot of these datasets contain huge amounts of information on only a limited number of subjects. A modern genotyped and imputed dataset can contain the genotypes of $\sim$ 10–100 million single nucleotide polymorphisms (SNPs), for a comparatively much smaller number of subjects ($\sim$ 10–500 thousand). The standard statistical techniques can provide biased estimates (*Baik and Silverstein*, 2006; *Lee et al.*, 2010) in the such high-dimensional regimes, where the number of features or covariates (p) is substantially larger than the number of observations (n), since the asymptotic properties of the estimators are profoundly different from the properties in low dimensional (p finite, $n \to \infty$) settings. Secondly, such large datasets impose enormous computational burden, and thus developing computationally efficient methods is of utmost importance. To address some of these problems with large datasets, we first focus on the issues arising from high-dimensionality in principal component analysis (PCA) and partial least squares (PLS), two of the most popular dimension reduction techniques used in high-dimensional biological data. Next, we focus on the problem of computational scalability in phenome-wide scale GWASs, and develop a single-variant test to address that.

1.1 Overview

1.1.1 Addressing the Consistency and Bias Problems of PCA in High-Dimensional Data

PCA is most commonly used to adjust for population stratification in GWASs (*Price et al.*, 2006), and to identify overall expression patterns in transcriptome analysis (*Storey et al.*, 2005). However, unlike in the low-dimensional setting, the sample eigenvalues and eigenvectors obtained from high-dimensional data are not consistent estimators of the population eigenvalues and eigenvectors (*Baik and Silverstein*, 2006; *Paul*, 2007; *Lee et al.*, 2010), and the predicted principal component (PC) scores can be systematically biased towards zero (*Lee et al.*, 2010). In the high-dimensional data literature, there has been extensive effort to study the convergence of sample eigenvalues, eigenvectors and PC scores (*Baik and Silverstein*, 2006; *Paul*, 2007; *Lee et al.*, 2010) under the spiked population model (*Johnstone*, 2001), which assumes that all eigenvalues are equal except for finitely many large ones. *Lee et al.* (2010) further proposed a bias-adjustment method for the predicted PC scores under this model. However, the equality assumption of the smaller eigenvalues in the spiked population model depends on the independence of the features, which may be violated in many real world datasets where the features are locally correlated. For example, in GWASs, nearby single nucleotide polymorphism (SNP) are highly correlated due to linkage disequilibrium. To accommodate such scenarios, we focus our research on the generalized spiked population model (*Bai and Yao*, 2012) which generalizes the spiked population model by dropping that equality assumption. In Chapter II, we systematically investigate the asymptotic behaviors of PCA under this model, and derive consistent estimators of the population eigenvalues, angles between the sample and population eigenvectors, correlation coefficients between the sample and population PC scores, and propose a method to adjust for the bias in the predicted PC scores. We demon-

strate the superior performance of our method by comparing it against the existing spiked population model-based method through extensive simulation studies and an application on the HapMap Phase III data (`http://hapmap.ncbi.nlm.nih.gov/`).

1.1.2 Addressing the Over-Fitting Problem of PLS in High-Dimensional Data

PLS, a closely related technique to PCA, is mostly used to fit regression models with high-dimensional predictors, due to its ability to simultaneously perform dimension reduction and model fitting. For example, in the genomics and transcriptomics literatures, it is applied to model different clinical outcomes on high-dimensional predictors such as sequence or gene expression data (*Boulesteix and Strimmer*, 2007; *Man*, 2004; *Huang et al.*, 2005; *Clementi et al.*, 1997). It is also used in the neurology literature to identify the functional patterns or anatomical regions of the brain that affect different neurological behaviors, or to model different chemical features based on information on a large number of metabolites (*Rubakhin et al.*, 2011; *Worley and Powers*, 2013) in the metabolomics literature. Even though PLS is an attractive tool to be applied when the number of predictors is large, or the predictors are correlated among themselves, it can suffer from an over-fitting problem in high-dimensional data, where the fitted outcomes become almost identical to the observed outcomes even when there is no or very little relation between the outcome and the predictors. This over-fitting problem has also been previously identified by other researchers (*Brereton and Lloyd*, 2014; *Gromski et al.*, 2015; *Lee et al.*, 2008; *He et al.*, 2017). However, no existing method can properly address this problem. In Chapter III, we first investigate the over-fitting problem, and propose a two-stage PLS (TPLS) method to address it, using theoretical results developed for high-dimensional PCA. Like high-dimensional PCA, we further notice a similar problem of biased predicted scores in high-dimensional PLS, and incorporate proper bias-adjustment procedures

in our method. We further evaluate a sparse variable selection-based method for high-dimensional PLS proposed by *Chun and Keleş* (2010), and compare it with our proposed method, in terms of addressing the over-fitting problem and prediction accuracy, using extensive simulated scenarios with various sparsity levels, and an application on the ADNI data.

1.1.3 Scalable and Accurate Single Variant Test for Unbalanced Case-Control GWAS and PheWAS

Next, we turn our attention to the computational burden of performing large-scale GWASs and phenome-wide association studies (PheWASs). Over the past decade, GWASs have successfully analyzed hundreds of diseases and traits and their associations with common genomic variations. Although asymptotic tests (score, Wald, likelihood ratio) are well-calibrated for a GWAS with binary phenotypes with balanced case-control ratios, it is a great challenge to develop single-variant tests that are scalable and accurate in the scale of a GWAS, when there are far fewer cases than controls. The asymptotic tests can provide substantially inflated type I error rates (*Ma et al.*, 2013) for rare (Minor allele frequency: MAF ≤ 0.01) and low frequency variants ($0.01 <$ MAF ≤ 0.05) in such situations. On the other hand, the Firth's test (*Firth*, 1993) is well-calibrated and robust for testing rare and low-frequency variants in unbalanced case-control studies. However, it is not computationally efficient as it needs to calculate the likelihood under the full model.

While it still remains difficult to analyze a GWAS for a binary phenotype with unbalanced case-control ratio, the researchers have proposed the PheWAS (*Denny et al.*, 2010) approach which is of substantially larger scale. PheWAS utilizes the detailed phenotypic information available from the EHR system in biobanks to construct a broad spectrum of human phenotypes or phenome, and allows researchers to exploit the cross-phenotype associations or pleiotropy (*Solovieff et al.*, 2013) phenomenon by

studying the impact of genetic variations across the phenome. Since genome-wide scale PheWASs attempts to perform genome-wide association analyses in 1000s of binary phenotypes, and most of them have unbalanced (case : control = 1 : 5) or often extremely unbalanced (case : control = 1 : 500) case-control ratios (see Figures 4.1,5.1), the existing single-variant tests are either not accurate (score test), or computationally so inefficient (Firth's test) that it is essentially impractical to apply them on a PheWAS. In Chapter IV, we propose a computationally efficient and accurate score test-based method (fastSPA) to test for binary phenotypes. Our method uses the saddlepoint approximation (*Daniels*, 1954) to provide a better approximation of the null distribution of the score statistic, than the traditionally used normal approximation. fastSPA is well-calibrated for controlling type I error rates and can adjust for other covariates, even for phenotypes with extremely unbalanced case-control ratios. In addition, it is ~ 100 times faster than the current gold standard Firth's test. For example, the projected computation time to test for 1500 phenotypes with 1 : 9 case-control ratio and 20000 sample size, across 10 million SNPs, for 20000 samples is ~ 400 CPU-days for our proposed test, compared to ~ 117 CPU-years for Firth's test.

1.1.4 Methods for Meta-Analyzing Multiple Unbalanced GWASs

As increasing number of association results from genome-wide scale PheWASs in different biobanks become available, meta-analyzing those association results is the logical next step to improve the power to detect novel genotype-phenotype associations. Because the binary phenotypes in biobank-based studies are mostly unbalanced in their case-control ratios, very few methods can provide well-calibrated tests for associations. For example, even though the Firth's test provides well-calibrated p values within individual studies, meta-analyzing them using the traditional Z-score-based method, which converts the individual p values into normal Z-scores and uses

their weighted sum as the final meta-analysis score, can results in conservative or anti-conservative type I error rates in such unbalanced scenarios (*Ma et al.*, 2013). In Chapter V, we show similar behavior of the Z-score-based method when meta-analyzing fastSPA-based p values. We further propose two meta-analysis strategies that can efficiently combine association results from these unbalanced GWASs. Our first method involves sharing an approximation of the distribution of the score test statistic from each study using cubic Hermite splines, and the second method involves sharing the overall genotype counts from each study. We demonstrate the performance of our meta-analysis methods in terms of controlling the type I errors using extensive simulation studies, and an application on the UK Biobank interim release data (*UK Biobank*, 2015).

CHAPTER II

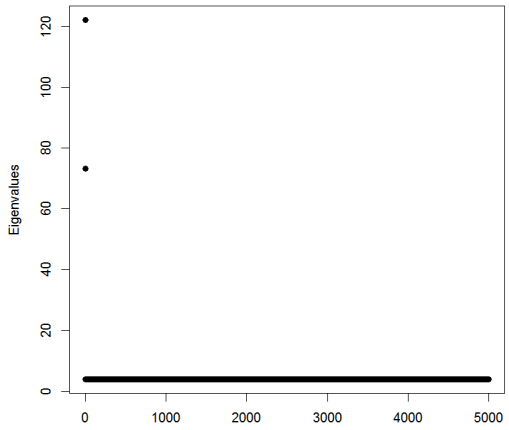
Asymptotic Properties of Principal Component Analysis and Shrinkage-Bias Adjustment under the Generalized Spiked Population Model

2.1 Introduction

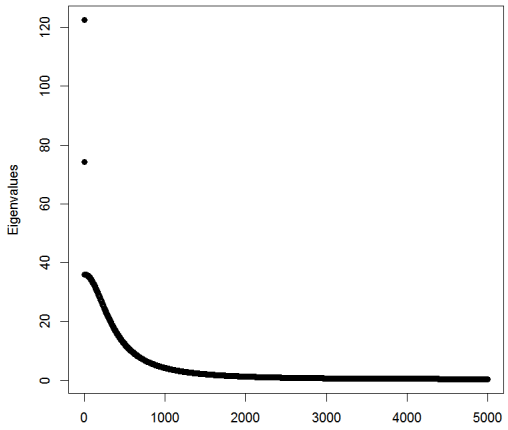
Principal component analysis (PCA) is a very popular tool for analyzing high-dimensional biomedical data, where the number of features (p) is often substantially larger than the number of observations (n). PCA is widely used to adjust for population stratification in genome-wide association studies (*Price et al.*, 2006) and to identify overall expression patterns in transcriptome analysis (*Storey et al.*, 2005). However, the asymptotic properties of PCA in high-dimensional data are profoundly different from the properties in low-dimensional (p finite, $n \to \infty$) settings. In high-dimensional settings, the sample eigenvalues and eigenvectors are not consistent estimators of the population eigenvalues and eigenvectors (*Johnstone and Lu*, 2009; *Paul*, 2007), and the predicted principal component (PC) scores based on the sample eigenvectors can be systematically biased toward zero (*Lee et al.*, 2010).

There has been extensive effort to investigate the asymptotic behaviors of PCA in high-dimensional settings. To provide a statistical framework for PCA in these settings, *Johnstone* (2001) introduced a spiked population model, which assumes that

all the eigenvalues are equal except for finitely many large ones (called the spikes). A spiked population covariance matrix is basically a finite rank perturbation of a scalar multiple of the identity matrix. A typical example of a spiked population with two spikes is shown in Figure 2.1(a). This two-spike eigenvalue structure arises if the population consists of three sub-populations which differ among themselves only through their means, and the features are largely independent with equal variances. Under this model, convergence of sample eigenvalues, eigenvectors and PC scores have been extensively studied (*Johnstone*, 2001; *Baik and Silverstein*, 2006; *Paul*, 2007; *Lee et al.*, 2010).



(a) Example of population eigenvalues under the spiked population model.



(b) Example of population eigenvalues in presence of an autoregressive within-group correlation structure. This clearly violates the assumptions of the spiked population model

Figure 2.1: Example of eigenvalues when the assumptions of the spiked population model are satisfied, and when they are violated.

In many biomedical data, however, the assumption of the equality of non-spiked eigenvalues can be violated due to the presence of local correlation among features. In genome-wide association studies, for example, the genetic variants are locally correlated due to linkage disequilibrium. In gene-expression data, since genes in the same pathway are often expressed together, their expression measurements are often correlated. These local correlations can cause substantial differences in non-spiked

eigenvalues. To illustrate this phenomenon, we obtained eigenvalues with an autoregressive within-group correlation structure rather than the independent structure of the previous example. Figure 2.1(b) shows that the equality assumption is clearly violated. Thus, if methods developed under the equality assumption are applied to these types of data, we will obtain biased results.

The generalized spiked population model (*Bai and Yao*, 2012) has been proposed to address this problem. The condition that the non-spikes have to be equal is removed in this generalization. In this model the set of population eigenvalues consists of finitely many large eigenvalues called the generalized spikes, which are well separated from infinitely many small eigenvalues. Although the generalized spiked population model has a great potential to provide more accurate inference in high-dimensional biomedical data, only limited literature is available on the asymptotic properties of PCA under this model and their application to real data. *Bai and Yao* (2012) and *Ding* (2015) provided results regarding convergence of eigenvalues and eigenvectors. However, their work remained largely theoretical. Moreover, to the best of our knowledge, no method has been developed for estimating the correlations between the sample and population PC scores, and adjusting biases in the predicted PC scores under the generalized spiked population model.

In this chapter, we systematically investigate the asymptotic behaviors of PCA under the generalized spiked population model, and develop methods to estimate the population eigenvalues and adjust for the bias in the predicted PC scores. We first propose two different approaches to consistently estimate the population eigenvalues, the angles between the sample and population eigenvectors, and the correlation coefficients between the sample and population PC scores. We compare these two methods and show the asymptotic equivalence of the estimators across them. Finally, we propose a method to reduce the bias in the predicted PC scores based on the estimated population eigenvalues.

2.2 Generalized Spiked Population Model

In order to formally define generalized spiked population model, we require the concept of spectral distribution. In random matrix literature, it is natural to associate a probability measure to the set of eigenvalues as the dimension (p) goes to ∞. More explicitly, if a Hermitian matrix Σ_p has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$, we can define the empirical spectral distribution (ESD) of Σ_p to be H_p based on the probability measure

$$dH_p(x) = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}(x),$$

where $\delta_{\lambda_i}(x)$ is unity when $x = \lambda_i$, and otherwise zero. Now, for a sequence $\{\Sigma_p\}$ of covariance matrices, if the corresponding sequence $\{H_p\}$ of ESDs converge weakly to a non-random probability distribution H as $p \to \infty$, then we define H as the limiting spectral distribution (LSD) of the sequence $\{\Sigma_p\}$.

The generalized spiked population model (*Bai and Yao*, 2012) is defined as follows. Suppose, H_p is the ESD corresponding to the population covariance matrix Σ_p and it converges weakly to a non-random probability distribution H. Let Γ_H be the support of H and $d(x, A) := \inf_{y \in A} |x - y|$ be the distance metric from a point x to a set A. Then the set of eigenvalues of Σ_p comprises of two subsets of eigenvalues $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_m$, and $\beta_{p,1} \geq \beta_{p,2} \geq \ldots \geq \beta_{p,p-m}$ where,

- Generalized spikes: $\exists \delta > 0$ such that $d(\alpha_i, \Gamma_H) > \delta$ for all $1 \leq i \leq m$. $\alpha_1, \ldots, \alpha_m$ are called the generalized spikes.

- Non-spikes: $\max_{1 \leq i \leq p-m} d(\beta_{p,i}, \Gamma_H) = \epsilon_p \to 0$. $\beta_{p,1} \geq \ldots \geq \beta_{p,p-m}$ are called the non-spikes.

It is obvious from the definition that the generalized spikes are measure zero points of the population LSD. For Johnstone's spiked population model (*Johnstone*, 2001), the population LSD is $H = \delta_{\{1\}}$, indicating $\Gamma_H = \{1\}$. From the definition above, all

eigenvalues larger than one are spikes. Hence, Johnstone's spiked population model is a special case of the generalized spiked population model.

Suppose that the population covariance matrix Σ_p has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, and the sample covariance matrix $S_p = X^T X/n$ has eigenvalues $d_1 \geq d_2 \geq \ldots \geq d_p$, where X is an $n \times p$ data matrix. Further, we will assume the following throughout this chapter:

Assumption 2.1. $p \to \infty, n \to \infty, p/n \to \gamma < \infty$.

Assumption 2.2. *The population eigenvalues follow the generalized spiked population (GSP) model with m generalized spikes. The population ESD H_p converges weakly to a non-random probability distribution H with support Γ_H. Moreover, the sequence $\{\|\Sigma_p\|\}$ of spectral norms is bounded. We will further assume that all the generalized spikes are larger than $\sup \Gamma_H$. Therefore, $\lambda_1, \ldots, \lambda_m$ are the generalized spikes, and the rest of the eigenvalues are considered as the non-spikes.*

Assumption 2.3. *The $n \times p$ data matrix $X = Y\Sigma_p^{1/2}$ where Y is an $n \times p$ random matrix with i.i.d. elements such that $E(Y_{ij}) = 0, E(|Y_{ij}|^2) = 1, E(|Y_{ij}|^4) < \infty$.*

Even though we will develop our estimation methods based on the asymptotic regime where $p/n \to \gamma < \infty$, we will discuss the applicability of our methods in ultra high-dimensional data where p is greatly larger than n in Section 2.4.6.

From the Marčenko–Pastur theorem (*Marčenko and Pastur*, 1967), the sample ESD F_p converges weakly to a non-random probability distribution F with support Γ_F. For $\alpha \notin \Gamma_H, \alpha \neq 0$ and $x > 0$, we define the following two functions

$$\psi(\alpha) := \alpha + \gamma\alpha \int \frac{\lambda dH(\lambda)}{\alpha - \lambda}, \quad f_F(x) := \frac{x}{1 + \gamma \int \frac{\tau dF(\tau)}{x - \tau}}. \tag{2.1}$$

The following result by *Bai and Yao* (2012) provides the almost sure limits of the sample eigenvalues corresponding to the population generalized spikes.

Result 2.1 (*Bai and Yao* (2012))**.** *Suppose Assumptions 2.1–2.3 hold. Let λ_k be a generalized spike of multiplicity one and the corresponding sample eigenvalue is d_k. Moreover, let ψ' denote the first derivative of the function ψ. Then,*

- *If $\psi'(\lambda_k) > 0$, then the sample eigenvalue d_k converges almost surely to $\psi(\lambda_k)$, i.e.*

$$|d_k - \psi(\lambda_k)| \xrightarrow{a.s.} 0.$$

- *If $\psi'(\lambda_k) \leq 0$, then let $(u_k, v_k) \subset (\sup \Gamma_H, \infty)$ be the maximal interval on which $\psi' > 0$. The sample eigenvalue d_k converges almost surely to $\psi(w)$ where w is a boundary of $[u_k, v_k]$ that is nearest to λ_k.*

Since $\psi'(\alpha)$ is a strictly increasing function for $\alpha > \sup \Gamma_H$, if a generalized spike λ_k is large enough such that $\psi'(\alpha) > 0$, according to Result 2.1 the corresponding sample eigenvalue will converge almost surely to $\psi(\lambda_k)$. However if the generalized spike lies close enough, i.e. $\psi'(\lambda_k) \leq 0$, to the set of non-spikes, then the convergence of the corresponding sample eigenvalue is given by the second part of the result. We will denote a generalized spike λ_k as a "distant spike" if $\psi'(\lambda_k) > 0$, otherwise we will call it a "close spike".

2.3 Consistent Estimation of the Generalized Spikes

The following theorem provides two different consistent estimators of the distant spikes.

Theorem 2.1. *Let λ_k be a distant spike of multiplicity one and the corresponding sample eigenvalue is d_k. If the Assumptions 2.1–2.3 hold, then,*

$$|\psi^{-1}(d_k) - \lambda_k| \xrightarrow{p} 0,$$

where ψ^{-1} is the left inverse of ψ. Also,

$$|f_F(d_k) - \lambda_k| \xrightarrow{p} 0.$$

This theorem shows that for any distant spike λ_k we have two consistent estimators $\psi^{-1}(d_k)$ and $f_F(d_k)$. Notice that the function f_F depends only on the sample LSD which can be approximated by the sample ESD. Thus, $f_F(d_k)$ can be approximated directly using the sample eigenvalues. More explicitly, $f_F(d_k)$ can be closely approximated as

$$f_F(d_k) \approx \frac{d_k}{1 + \frac{\gamma}{p-m} \sum_{i=m+1}^{p} \frac{d_i}{d_k - d_i}}.$$

In contrast, the ψ function depends on the population LSD which is unknown. We can estimate the ψ function using the algorithm described in Section 2.5 and then find the inverse function ψ^{-1} using a Newton-Raphson type algorithm.

2.4 Consistent Estimators of the Asymptotic Shrinkage in the Predicted PC Scores

In this section, we investigate the convergence of sample eigenvectors, PC scores, and shrinkage factors in predicting the PC scores. Let e_i and E_i to be the i^{th} sample and population eigenvectors, respectively. In addition to Assumptions 2.1–2.3, we further assume that the distant spikes are of multiplicity one. This assumption is to restrict the dimension of the corresponding eigenspaces to one, as otherwise the angle between sample and population eigenvectors, or shrinkage in predicted PC scores cannot be well defined.

2.4.1 Angle between Sample and Population Eigenvectors

We first present the following theorem on the convergence of the quadratic forms of the sample eigenvectors.

Theorem 2.2. *Let λ_k be a distant spike of multiplicity one, and the Assumptions 2.1–2.3 hold. Consider the following quadratic form*

$$\hat{\eta_k} = s_1^T e_k e_k^T s_2,$$

where s_1 and s_2 are non-random vectors with uniformly bounded norm for all p. Then,

$$|\hat{\eta_k} - \eta_k| \xrightarrow{a.s.} 0,$$

where

$$\eta_k = \frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)} s_1^T E_k E_k^T s_2$$

Mestre (2008b) showed similar asymptotic properties of the quadratic forms under the assumption that the number of spikes increases with the dimension. Theorem 2.2 shows the convergence of the angle between sample and population eigenvectors. Suppose $s_1 = s_2 = E_k$, and then,

$$\hat{\eta_k} = E_k^T e_k e_k^T E_k = \langle e_k, E_k \rangle^2, \quad \eta_k = \frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)}.$$

Combining them, we can show

$$\left| \langle e_k, E_k \rangle^2 - \frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)} \right| \xrightarrow{a.s.} 0. \tag{2.2}$$

Therefore, $\{\lambda_k \psi'(\lambda_k)/\psi(\lambda_k)\}^{1/2}$ is a consistent estimator of the cosine of the angle, i.e. the absolute value of the inner product, between the k^{th} sample and population eigenvectors. In order to obtain this estimator we first need to estimate the ψ function

using the algorithm described in Section 2.5.

The following result by *Ding* (2015) provides another consistent estimator for the angle between the k^{th} sample and population eigenvectors. The proof of the asymptotic equivalence of these two estimators is given in Appendix A.

Result 2.2. *Let λ_k be a distant spike of multiplicity one, and d_k be the corresponding sample eigenvalue. Suppose that Assumptions 2.1–2.3 hold. Define,*

$$g_F(x) := \left[1 + \gamma f_F(x) \int \frac{\tau dF(\tau)}{(x-\tau)^2}\right]^{-1}.$$

Then,

$$|\langle e_k, E_k \rangle|^2 - g_F(d_k)| \xrightarrow{p} 0.$$

Hence $g_F(d_k)^{1/2}$ also works as a consistent estimator of $|\langle e_k, E_k \rangle|$. Since the function g_F depends only on sample LSD, it can be approximated directly using sample eigenvalues. More explicitly, if there are m spikes in the population, the function g_F can be closely approximated as

$$g_F(d_k) \approx \left[1 + \frac{\gamma f_F(d_k)}{p-m} \sum_{i=m+1}^{p} \frac{d_i}{(d_k - d_i)^2}\right]^{-1}.$$

The above equation can be used to estimate the angle between the sample and population eigenvectors.

2.4.2 Correlation between Sample and Population PC Scores

The sample and population PC scores are the projections of the data on the sample and population eigenvectors respectively. The correlation between them can be perceived as a measure of accuracy of the PCA. The squared correlation can also be interpreted as the proportion of variance in the population PC scores that can be explained by corresponding sample PC scores. The following theorem provides the

consistent estimators of the correlation between the sample and population PC scores corresponding to a distant spike.

Theorem 2.3. *Suppose λ_k is a distant spike of multiplicity one, d_k is the corresponding sample eigenvalue, and the Assumptions 2.1–2.3 hold. Let the normalized k^{th} population PC score is $P_k = XE_k/(n\lambda_k)^{1/2}$ and the normalized k^{th} sample PC score is $p_k = Xe_k/(nd_k)^{1/2}$. Then,*

$$\left|\langle P_k, p_k\rangle^2 - \psi'(\lambda_k)\right| \xrightarrow{p} 0,$$

and,

$$\left|\langle P_k, p_k\rangle^2 - \frac{d_k g_F(d_k)}{f_F(d_k)}\right| \xrightarrow{p} 0,$$

where the function g_F is as defined in Result 2.2.

Since P_k and p_k are normalized random vectors, the absolute value of the inner product $\langle P_k, p_k\rangle$ is identical to the absolute value of their correlation coefficient. Since correlation is scale invariant, this is also the correlation between k^{th} sample and population PC scores. Therefore we can consider both $\psi'(\lambda_k)^{1/2}$ and $\{d_k g_F(d_k)/f_F(d_k)\}^{1/2}$ to be consistent estimators of the correlation between the k^{th} sample and population PC scores.

2.4.3 Asymptotic Shrinkage Factor

Suppose λ_k is a distant spike. Let the k^{th} sample PC score for the j^{th} observation x_j be $p_{kj} = x_j^T e_k$, and the k^{th} predicted PC score for a new observation x_{new} be $q_k = x_{new}^T e_k$. Then the quantity $\rho_k = \lim_{p\to\infty}\{E(q_k^2)/E(p_{kj}^2)\}^{1/2}$ describes the asymptotic shrinkage in the k^{th} predicted PC score for a new observation. As both p_{kj} and q_k are centered, i.e. $E(p_{kj}) = E(q_k) = 0$, ρ_k represents the limiting ratio of the standard deviations of the predicted PC scores and the sample PC scores. Therefore, if we can estimate ρ_k, then the shrinkage bias in the k^{th} predicted PC scores can be easily

adjusted by rescaling the predicted scores by the factor ρ_k^{-1}. The following theorem provides the consistent estimator of the asymptotic shrinkage factor ρ_k.

Theorem 2.4. *Suppose λ_k is a distant spike of multiplicity one, d_k is the corresponding sample eigenvalue, and the Assumptions 2.1–2.3 hold. Let p_{kj} and q_k be as defined above. Then,*

$$\left| \sqrt{\frac{E(q_k^2)}{E(p_{kj}^2)}} - \frac{\lambda_k}{d_k} \right| \xrightarrow{p} 0.$$

This is a surprising result in which the asymptotic shrinkage factor is expressed as a simple ratio of the population and sample eigenvalues. Recall that we already constructed the consistent estimators for population eigenvalues in the previous sections. Using these results, the asymptotic shrinkage factor ρ_k can be consistently estimated by $\hat{\lambda}_k/d_k$ where $\hat{\lambda}_k$ is any consistent estimator of λ_k.

2.4.4 Comparison between the Two Different Estimators

For each of the quantities discussed above, we proposed two asymptotically equivalent estimators. In terms of practical applications they have their own advantages and disadvantages. One of them can be approximated directly based only on the sample eigenvalues, while the other one requires to estimate the LSD of the population eigenvalues to obtain the ψ function. For ease of discourse we will call the former "d-estimator" and the later "λ-estimator". If the number of spikes is known, estimating the d-estimator is computationally more efficient than estimating the λ-estimator as it does not involve estimating the population LSD. However, by estimating the population LSD the λ-estimation procedure can verify whether an estimated eigenvalue is actually a distant spike by checking if $\psi' > 0$. Thus it can be used to estimate the number of distant spikes when it is unknown (see Section 2.5). On the other hand, the d-estimation procedure provides no information on the population LSD and thus cannot distinguish among distant spikes, close spikes and non-spikes. To summarize,

when the number of spikes is known or we only want to estimate few of the largest eigenvalues which are known to be distant spikes, then the d-estimation procedure has the advantage of a faster computation, while the λ-estimation procedure is more useful when the number of spikes is unknown or the distribution of the non-spikes is of interest.

2.4.5 Comparison between the Generalized Spiked Population (GSP) Model and the Spiked Population (SP) Model

As mentioned before, the SP model (*Johnstone*, 2001) is a special case of the GSP model. It is easy to verify that when the population eigenvalues follow the SP model, our consistent estimators for the spiked eigenvalues, the angles between the eigenvectors, the correlation coefficients between the PC scores and the shrinkage factors conform to the consistent estimators derived by *Baik and Silverstein* (2006); *Paul* (2007); *Lee et al.* (2010). For an SP model where all the non-spikes are equal to one, the LSD H is a degenerate distribution at one, and

$$\psi(\alpha) = \alpha\left(1 + \frac{\gamma}{\alpha - 1}\right); \quad \psi'(\alpha) = 1 - \frac{\gamma}{(\alpha-1)^2}.$$

Now, $\psi'(\alpha) > 0$ if and only if $\alpha > 1 + \gamma^{1/2}$. If $\alpha > 1 + \gamma^{1/2}$ and d is the corresponding sample eigenvalue, then the consistent estimator of α is given by $\psi^{-1}(d)$, and

$$\frac{\alpha\psi'(\alpha)}{\psi(\alpha)} = \frac{1 - \frac{\gamma}{(\alpha-1)^2}}{1 + \frac{\gamma}{\alpha-1}}; \quad \frac{\alpha}{\psi(\alpha)} = \frac{\alpha - 1}{\alpha + \gamma - 1},$$

which show that all our results match with the results from *Lee et al.* (2010).

It is of interest to investigate how closely methods developed under the SP model can approximate the consistent estimators for the distant spikes when the population eigenvalues actually follow a GSP model. Suppose the population eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ follow the GSP model with m distant spikes. The sample eigenvalues

are $d_1 \geq d_2 \geq \ldots \geq d_p$. Let λ_k be a distant spike with multiplicity one, and the corresponding sample eigenvalue is d_k. Then according to Result 2.1, $d_k \to \psi(\lambda_k)$ almost surely. From the definition of ψ,

$$\psi(\lambda_k) = \lambda_k \left(1 + \gamma \int \frac{\lambda dH(\lambda)}{\lambda_k - \lambda}\right) = \lambda_k + \gamma \int \frac{\lambda dH(\lambda)}{1 - \lambda/\lambda_k}.$$

If H is almost degenerate, i.e., the non-spikes are nearly identical, then,

$$\psi(\lambda_k) \approx \lambda_k + \frac{\gamma \bar{\lambda}}{1 - \bar{\lambda}/\lambda_k}, \tag{2.3}$$

where $\bar{\lambda} = \int \lambda dH(\lambda)$ is the mean of the population LSD which can be closely approximated by the mean of the non-spikes. On the other hand, if the spike λ_k is very large compared to all the non-spikes such that $\lambda/\lambda_k \approx 0$ for any $\lambda \in \Gamma_H$, then

$$\psi(\lambda_k) \approx \lambda_k + \gamma \bar{\lambda}. \tag{2.4}$$

Now, suppose instead of using the GSP assumption, we use the SP assumption to estimate the distant spikes. We assume that under the SP model the population covariance matrix is scaled by a factor ζ and the population eigenvalues are $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_m > \zeta = \zeta = \ldots \zeta$. If β_k is the population eigenvalue corresponding to d_k, then $d_k \to \psi(\beta_k)$ almost surely where,

$$\begin{aligned} \psi(\beta_k) &= \beta_k \left(1 + \gamma \frac{\zeta}{\beta_k - \zeta}\right) \\ &= \beta_k + \frac{\gamma \zeta}{1 - \zeta/\beta_k}. \end{aligned}$$

Here ζ is estimated as the mean of the non-spikes as they are all assumed to be equal to ζ. Notice that this expression is approximately equal to the expression in (2.3) with $\beta_k = \lambda_k$ and $\zeta = \bar{\lambda}$. Therefore, the asymptotic limit of d_k under both the GSP

and the SP model are approximately equal when the non-spikes are nearly identical. On the other hand, when the spike β_k is very large compared to all the non-spikes such that $\zeta/\beta_k \approx 0$, then $\psi(\beta_k) \approx \beta_k + \gamma\zeta$. In this case also, the asymptotic limit of d_k under both the GSP and the SP model are approximately equal with $\beta_k = \lambda_k$ and $\zeta = \bar{\lambda}$. Therefore if a generalized spike is very far away from the support of the population LSD, then the estimate of the spike based on an SP model will closely approximate the estimate based on a GSP model. However the SP model will provide potentially biased estimates if the non-spikes are not similar and the ratio between the largest non-spike and the spike of interest is substantially larger than zero.

2.4.6 Comparison with Ultra High-Dimensional Regime-Based Results When p/n is Large

Our methods are developed under the high-dimensional regime $p/n \to \gamma < \infty$, and does not theoretically warrant it to be applied in UHD regime where $p/n \to \infty$. However, often times in real world applications, we only have data with large p and large n, but the relative rate of their asymptotic divergence is unknown. Therefore, we do not know whether the true asymptotic regime is high-dimensional ($p/n \to \gamma < \infty$) or ultra high-dimensional ($p/n \to \infty$). Suppose that the true asymptotic regime is high-dimensional with γ finite but large compared to n, and the eigenvalues follow the GSP model. In such situations, we can either correctly assume the high-dimensional regime and apply the results discussed in this chapter, or we can falsely assume the ultra high-dimensional asymptotic regime and employ the theoretical results derived under this regime (*Lee et al.*, 2014b). In this section, we will investigate whether it is prudent to assume the UHD regime in such situations. In other words, we will try to answer how large γ can be considered to be diverging to infinity for practical applications.

We first show that for large enough γ, the theoretical results based on the falsely

assumed UHD regime become nearly identical to the results under the correctly assumed GSP (under high-dimensional regime) model. The UHD-based results presented in *Lee et al.* (2014b) require weaker conditions for the non-spiked eigenvalues than those for the spiked population model. Instead of assuming that they are the same, it assumes certain conditions on the moments of the non-spiked eigenvalues. Since the population LSD has a finite support and all of its central moments are finite, the condition on their moments, i.e. condition 2 in *Lee et al.* (2014b), is satisfied with an additional assumption that $n^3/p^2 = o(1)$. Without loss of generality, we assume that the mean of the non-spikes is unity. Then, under the UHD regime,

$$\frac{d}{\lambda} \xrightarrow{p} \frac{\gamma}{\lambda} + 1 \quad \text{when } \lambda \geq O(\gamma); \quad \frac{d}{\gamma} \xrightarrow{p} 1 \quad \text{when } \lambda = o(\gamma), \tag{2.5}$$

where λ and d are a spiked population eigenvalue and its corresponding sample eigenvalue, respectively. Here $\lambda \geq O(\gamma)$ means λ/γ is bounded away from zero, and $\lambda = o(\gamma)$ means $\lambda/\gamma \to 0$. They also showed the convergence of sample eigenvectors and PC scores.

Alternatively, under the GSP model, $d \to \psi(\lambda)$ when λ is a distance spike. From Theorem 2.1, a distant spike λ must satisfy

$$1 - \gamma \int \frac{x^2 dH(x)}{(\lambda - x)^2} > 0,$$

where H is the population LSD. Since $f_\lambda(x) = x^2(\lambda - x)^{-2}$ is a continuous function for $\lambda > \sup \Gamma_H$ and $x \in \Gamma_H$, where Γ_H is the support of H, there exists $x^* \in (\inf \Gamma_H, \sup \Gamma_H)$ such that $\int x^2(\lambda - x)^{-2} dH(x) = x^{*2} (\lambda - x^*)^{-2}$. Then,

$$1 - \gamma \frac{x^{*2}}{(\lambda - x^*)^2} > 0,$$

which implies $\lambda > x^* + x^*\sqrt{\gamma}$. Thus, for any $\lambda > x^* + x^*\sqrt{\gamma}$, $d/\lambda - \psi(\lambda)/\lambda$ converges

to zero.

Now, under the true asymptotic regime (high-dimensional) λ and γ are both finite and non-zero, and thus $\lambda = O(\gamma)$. However, under the falsely assumed UHD regime, one can further assume $\lambda \geq O(\gamma)$ or $\lambda = o(\gamma)$ depending on whether λ is large or small compared to γ. If one assumes $\lambda \geq O(\gamma)$, then the difference between the convergence of d/λ from the two models is

$$\frac{\psi(\lambda)}{\lambda} - \frac{\gamma}{\lambda} - 1 = \frac{\gamma}{\lambda}\left(\int \frac{x dH(x)}{1 - x/\lambda} - 1\right). \tag{2.6}$$

Since $\gamma/\lambda = O(1)$ and $\int x\left(1 - x/\lambda\right)^{-1} dH(x) - 1 = O\left(\lambda^{-1}\right)$ as the mean of the non-spikes is unity, (2.6) becomes almost identical to zero when λ is sufficiently large.

Now, suppose one assumes $\lambda = o(\gamma)$. Let $\lambda \simeq a + b\gamma^k$ for some finite a, b and $1/2 \leq k < 1$. Then, the difference between our result and the UHD result is

$$\left|\frac{\psi(\lambda)}{\gamma} - 1\right| = \left|\frac{\lambda}{\gamma} + \lambda \int \frac{x dH(x)}{\lambda - x} - 1\right| = O(\gamma^{k-1}). \tag{2.7}$$

Thus, in this case also (2.7) becomes almost identical to zero when γ is sufficiently large. We can also show the similar results for eigenvectors and PC scores.

Although both GSP and UHD eventually provide nearly identical results when γ is sufficiently large, the GSP model can provide substantially better estimates. The difference can be large when λ is small compare to γ, i.e $k < 1$, since the difference in (2.7) is of the order $O\left(\gamma^{k-1}\right)$. The difference will be at least as large as $O\left(1/\sqrt{\gamma}\right)$ in such cases. In simulation studies, we show this numerically. Therefore, in the scenario where γ is large compared to n, our suggestion would be to use the UHD method only when we apriori know that the spike is very large compared to γ. Otherwise, our GSP model based methods will provide better estimates.

2.5 Estimation of the Population Limiting Spectral Distribution

The λ-estimators rely on ψ, that is a function of the unknown population LSD H. To use the λ-estimators, it is thus required to estimate H. Using the Stieltjes transformation and the Marčenko–Pastur theorem, *El Karoui* (2008) developed a general algorithm to estimate the population LSD from the sample ESD, F_p. We propose to use Karoui's method to estimate the population LSD H and then use it to estimate ψ.

2.5.1 Karoui's Algorithm

Suppose v_{F_p} is the Stieltjes transformation of the set of eigenvalues in the sample covariance matrix in which

$$v_{F_p}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i - z}$$

for any $z \in \mathbb{C}^+, \mathbb{C}^+ = \{x \in \mathbb{C}, Im(x) > 0\}$. According to the Marčenko–Pastur theorem (*Marčenko and Pastur*, 1967), when Assumptions 2.1–2.3 hold, v_{F_p} converges pointwise almost surely to a non-random limit v_F, which uniquely satisfies the following equation

$$v_F(z) = -\left(z - \gamma\int\frac{\lambda dH(\lambda)}{1+\lambda v_F(z)}\right)^{-1}.$$

Karoui's method first calculates v_{F_p} for a grid of values $\{z_j\}_{j=1}^{J}$, and then finds $\hat{H}$ as a solution to minimize the following objective function

$$\hat{H} = \arg_H \min L\left(\left\{\frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n}\int\frac{\lambda dH(\lambda)}{1+\lambda v_{F_p}(z_j)}\right\}_{j=1}^{J}\right),$$

where L is any pre-defined convex loss function. In order to approximate the integral inside of the loss function, the algorithm discretizes H in the following way,

$$dH(\lambda) \simeq \sum_{k=1}^{K} w_k \delta_{t_k}(\lambda),$$

where $\delta_{t_k}(\lambda) = 1$ if $\lambda = t_k$ and 0 otherwise, $\sum_{k=1}^{K} w_k = 1$ with $w_k > 0$ for all k, and $\{t_k\}_{k=1}^{K}$ is a grid of points on the support of H. This is basically approximating H by a discrete distribution with support $\{t_k\}_{k=1}^{K}$. Then the integral is approximated by

$$\int \frac{\lambda dH(\lambda)}{1 + \lambda v_F(z)} \simeq \sum_{k=1}^{K} w_k \frac{t_k}{1 + t_k v_{F_p}(z_j)},$$

and the minimization problem transforms into,

$$\hat{H} = \arg_H \min L \left(\left\{ \frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n} \sum_{k=1}^{K} w_k \frac{t_k}{1 + t_k v_{F_p}(z_j)} \right\}_{j=1}^{J} \right). \tag{2.8}$$

El Karoui (2008) has shown the weak convergence of $\hat{H}$ to H, i.e $\hat{H} \to H$.

Some examples of the convex loss function L can be,

- $L_\infty \left(\{e_j\}_{j=1}^{J} \right) = \max_j \max \{|Re(e_j)|, |Im(e_j)|\}$
- $L_1 \left(\{e_j\}_{j=1}^{J} \right) = \sum_{j=1}^{J} |e_j|$
- $L_2 \left(\{e_j\}_{j=1}^{J} \right) = \sum_{j=1}^{J} |e_j|^2$

For the convex loss functions described above, the estimation of H in (2.8) reduces to a convex optimization problem (*Boyd and Vandenberghe*, 2004). Karoui also provided a translation of this problem into a linear programming problem when L_∞ loss function is used. Further details can be found in *El Karoui* (2008).

2.5.2 Implementing Karoui's Algorithm When the Number of Spikes is Known

Since the generalized spikes fall outside the support of the population LSD, Karoui's algorithm cannot be directly applied to estimate the spikes. Furthermore, *Bai and Silverstein* (1998) showed that the probability of a sample eigenvalue falling outside the support of the sample LSD will go to zero as p increases, which implies that the sample eigenvalues corresponding to the population generalized spikes will be measure zero points in the sample LSD. Since the spikes behave like measure zero points (or outliers) when we are concerned about estimating the population LSD, we can exclude the sample eigenvalues corresponding to the population generalized spikes while calculating v_{F_p} and that will lead to a more robust estimation of H. Therefore, we will apply Karoui's algorithm in the following way,

1. Suppose the population covariance matrix possesses m generalized spikes. We exclude the top m sample eigenvalues while calculating v_{F_p},

$$v_{F_p}(z) = \frac{1}{n-m}\sum_{i=1}^{n}\frac{1}{d_i - z}.$$

2. Apply Karoui's algorithm to obtain $\hat{H}$. Further, if it is reasonable to assume that the true population LSD is a continuous or piecewise continuous distribution function, suitable kernel smoothing algorithm can be used on $\hat{H}$ to obtain a more continuous approximation of H.

3. The quantiles of $\hat{H}$ can be considered as the estimators of the non-spikes.

4. Suppose, $\hat{\lambda}_{m+1}, \hat{\lambda}_{m+2}, \ldots, \hat{\lambda}_p$ are the estimated non-spikes. Then the ψ function is estimated by,

$$\hat{\psi}(\alpha) = \alpha + \frac{\gamma\alpha}{p-m}\sum_{i=m+1}^{p}\frac{\hat{\lambda}_i}{\alpha - \hat{\lambda}_i}.$$

Due to the weak convergence $\hat{H} \to H$, $\hat{\psi}$ will also converge to ψ point-wise. Thus, all the estimates provided in Section 2.3 and 2.4 will still be consistent if we replace ψ with $\hat{\psi}$.

2.5.3 Estimating the Number of Spikes

Our application of Karoui's algorithm to the GSP model depends on the number of spikes m, which is usually unknown. If we have some knowledge of the underlying structure of the data, we can use it to estimate m roughly. Suppose we know that the data are coming from a mixture of K subpopulations, and within each subpopulation the observations are i.i.d. $N(\mu_k, \Sigma)$, where μ_k represents the mean for the k^{th} subpopulation, and Σ is the common within-group population covariance matrix. Then, as the spikes represent the between group differences, the number of spikes should be the same as the rank of the between group covariance matrix which is $(K-1)$. However in real data, it is often hard to accurately assess the number of such homogeneous subpopulations. In those cases we can use the following algorithm to estimate m.

1. Start with a reasonable finite upper bound m_{max} of the number of spikes. The upper bound can be selected based on prior information on the subpopulations, or by examining the sample eigenvalues. Set $m = m_{max}$.

2. Use Karoui's algorithm to estimate the population LSD and the non-spikes. Suppose the estimated non-spikes are $\hat{\lambda}_{m+1} \geq \hat{\lambda}_{m+2} \geq \ldots \geq \hat{\lambda}_p$, and the ψ function is estimated by,

$$\hat{\psi}(\alpha) = \alpha + \frac{\gamma\alpha}{p-m} \sum_{i=m+1}^{p} \frac{\hat{\lambda}_i}{\alpha - \hat{\lambda}_i}.$$

3. Find $S_\psi > \lambda_{m+1}$ using Newton-Raphson algorithm such that

$$\hat{\psi}'(S_\psi) = 1 - \frac{\gamma}{p-m} \sum_{i=m+1}^{p} \left(\frac{\hat{\lambda}_i}{S_\psi - \hat{\lambda}_i} \right)^2 = 0.$$

4. Since any distant spike must be larger than S_ψ, and $\hat{\psi}, \hat{\psi}'$ are both continuous and strictly increasing functions on (S_ψ, ∞), the equation $\hat{\psi}(\lambda) - d_k = 0$ has a root in (S_ψ, ∞) if and only if $\hat{\psi}(S_\psi) - d_k < 0$. Therefore, find the smallest index i^* in $1, 2, \ldots, m$ such that $d_{i^*} \leq \hat{\psi}(S_\psi)$. If all $d_1, d_2, \ldots, d_m$ are larger than $\hat{\psi}(S_\psi)$ then stop and select m as the number of distant spikes. Otherwise, set $m = i^* - 1$ and repeat steps 2–4.

Note that the close spikes occur so close to the support of the population LSD that they cannot be distinguished separately from the non-spikes when the number of spikes is unknown.

The selection of m_{max} is subjective. It can be selected based on the prior knowledge on the number of subpopulations, or by investigating the sample eigenvalues. In real data applications, we are usually interested in only a few large eigenvalues. In such situations, m_{max} can also be selected to be slightly larger than the number of eigenvalues we are interested in. As seen from our simulation studies, this spike selection algorithm can overestimate the number of spikes if the upper bound m_{max} is too large or underestimate the number of spikes if there are close spikes present (Table B.1). However, as long as m_{max} is finite compared to n and p, the estimation of the true population distant spikes will still remain consistent.

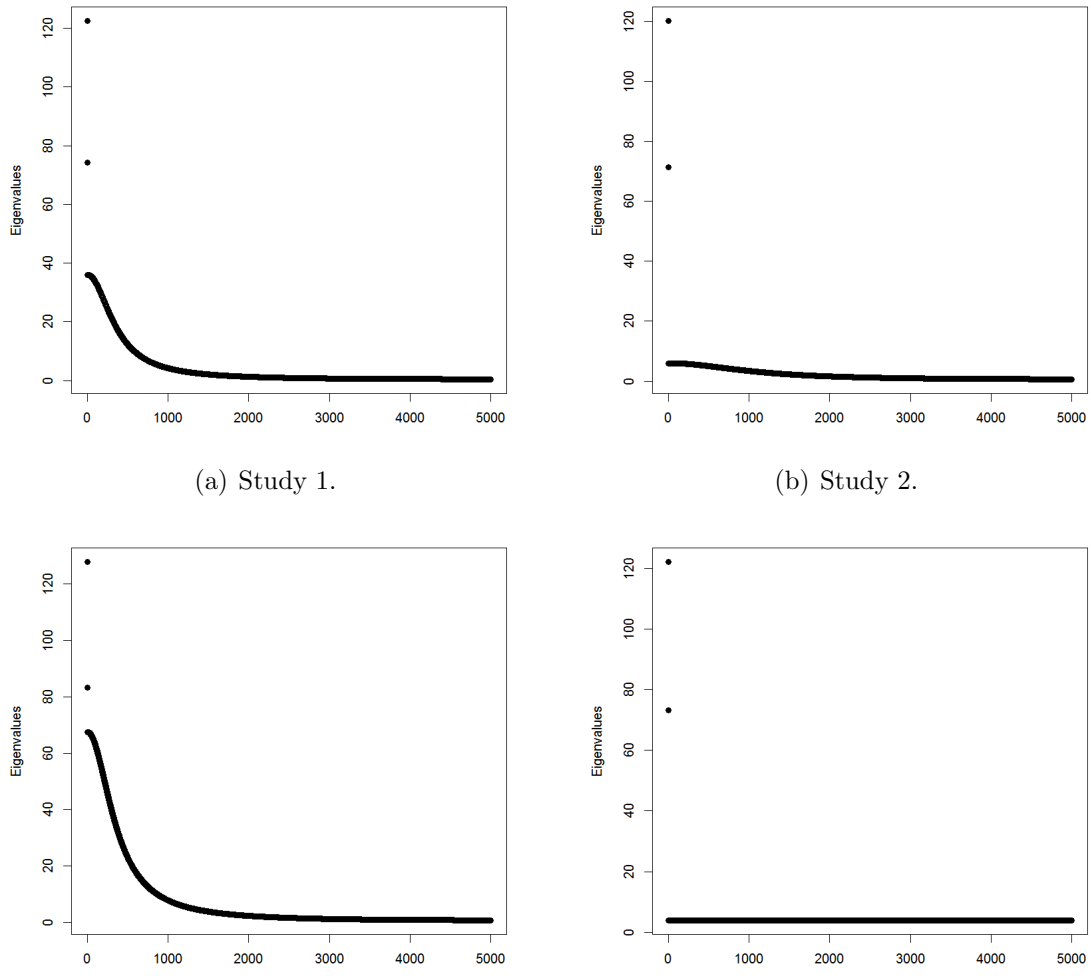
2.6 Simulation Studies and Real Data Example

2.6.1 Simulation Studies: Compare GSP and SP-Based Methods

In this section we will present simulation studies of four different scenarios to compare the performances of the proposed GSP-based methods and the existing SP-based method proposed by *Lee et al.* (2010). For each study, we simulated a training dataset with $n = 500$ individuals and $p = 5000$ features. The data were generated from three subpopulations with sample sizes 100, 150 and 250. For each subpopulation we first selected a mean vector μ_i by drawing its elements randomly with replacement from $\{-0.3, 0, 0.3\}$. Then samples in the i^{th} subpopulation were drawn from $N_p(\mu_i, V)$ where V is the AR(1) covariance matrix with variance σ^2 and autocorrelation ρ. The (σ^2, ρ) pairs used for the four studies were $(4, 0.8), (1, 0.7), (7.5, 0.8)$ and $(4, 0)$. The population eigenvalue plots for all the studies are shown in Figure 2.2.

We also generated test datasets for each study with the same settings as the training datasets. Then we applied our GSP-based methods and the existing SP-based method to estimate the population spikes, the angles between the sample and population eigenvectors, the correlations between the sample and population PC scores and the asymptotic shrinkage factors. For all of the studies, we used the upper bound $m_{max} = 5$ to estimate the number of distant spikes using the algorithm described in Section 2.5.3. We simulated each study 200 times to calculate the empirical biases and standard errors of the estimates. The results are presented in Table 2.1.

It is clear from Table 2.1 that for Study 1, 2 and 3 our methods reduced the bias in all the estimates while having similar standard errors as the existing method. The positive empirical biases in all the SP estimates suggest that the SP method tends to overestimate all the quantities. In Study 4, since the underlying population satisfied the SP assumption, all methods provided very similar and almost unbiased estimates ($< 1\%$). The results also verify that the λ-estimates and d-estimates are



(a) Study 1.

(b) Study 2.

(c) Study 3.

(d) Study 4.

Figure 2.2: Eigenvalue structures in simulation studies comparing GSP-based and SP-based methods.

Settings		**Method**	**Eigenvalue**		**Angle**		**Correlation**		**Shrinkage**	
No.			**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
1	$n = 500$	SP	5.27	18.27	6.52	34.07	3.83	23.33	5.32	17.88
	$p = 5000$		(2.37)	(3.11)	(0.32)	(0.60)	(0.03)	(0.08)	(0.60)	(1.06)
	$\sigma^2 = 4$	λ-GSP	0.43	0.95	0.53	3.28	0.33	2.79	0.47	0.58
	$\rho = 0.8$		(2.67)	(5.27)	(0.77)	(6.26)	(0.31)	(4.69)	(0.92)	(3.31)
		d-GSP	0.47	0.69	0.47	2.48	0.24	2.11	0.51	0.31
			(2.67)	(5.45)	(0.77)	(6.70)	(0.31)	(5.07)	(0.92)	(3.51)
2	$n = 500$	SP	0.10	0.46	0.16	0.44	0.08	0.24	0.18	0.39
	$p = 5000$		(0.90)	(1.27)	(0.04)	(0.08)	(0.001)	(0.003)	(0.08)	(0.16)
	$\sigma^2 = 1$	λ-GSP	-0.04	0.04	0.01	0.004	0.01	0.01	0.03	-0.03
	$\rho = 0.7$		(0.90)	(1.28)	(0.04)	(0.10)	(0.03)	(0.01)	(0.08)	(0.18)
		d-GSP	-0.004	0.10	0.03	0.03	0.004	0.01	0.07	0.03
			(0.90)	(1.28)	(0.04)	(0.10)	(0.03)	(0.01)	(0.08)	(0.18)
3	$n = 500$	SP	25.68	-	64.06	-	46.50	-	26.41	-
	$p = 5000$		(2.54)		(0.52)		(0.07)		(0.90)	
	$\sigma^2 = 7.5$	λ-GSP	2.92	-	12.62	-	10.95	-	3.47	-
	$\rho = 0.8$		(5.7)		(11.90)		(10.13)		(4.20)	
		d-GSP	2.45	-	12.25	-	10.87	-	3.00	-
			(5.74)		(10.52)		(8.58)		(4.24)	
4	$n = 500$	SP	0.05	-0.26	0.06	-0.06	0.03	0.05	0.07	-0.22
	$p = 5000$		(1.58)	(2.35)	(0.23)	(0.53)	(0.02)	(0.08)	(0.43)	(0.90)
	$\sigma^2 = 4$	λ-GSP	0.03	-0.35	0.02	-0.18	0.01	-0.02	0.04	-0.31
	$\rho = 0$		(1.58)	(2.35)	(0.24)	(0.54)	(0.02)	(0.09)	(0.43)	(0.91)
		d-GSP	0.16	-0.12	0.10	-0.03	0.01	0.02	0.18	-0.08
			(1.58)	(2.35)	(0.23)	(0.53)	(0.02)	(0.09)	(0.42)	(0.90)

Table 2.1: Simulation results for GSP-based and SP-based methods. All methods were applied to estimate the population eigenvalues, cosine of the angles between sample and population eigenvectors, correlations between sample and population PC scores, and the asymptotic shrinkage factors. Each cell has empirical bias (%) with coefficients of variations (%) in parenthesis.

asymptotically equivalent. The performances of the λ-estimates and the d-estimates are nearly identical in all the simulation studies.

In Study 1, the ratio of the largest non-spike with the two spikes are 0.29 and 0.48, which are substantially larger than zero. Thus according to the discussion in Section 2.4 the SP model does not closely approximate the GSP model. The results support this assertion as the SP model-based estimates are highly biased whereas the estimates based on our methods have very little empirical bias. On the other hand, in Study 2 the largest non-spike is very small compared to the smallest spike (ratio 0.08). Thus the estimates based on the SP model closely approximate the estimates based on the GSP model, and we find very little empirical bias ($< 1\%$) in all of the SP model-based estimates. In Study 3, even though there were two spikes present, only the largest population eigenvalue was a distant spike. So we presented only the estimates corresponding to the largest population eigenvalue. Since the ratio of the largest non-spike and the largest spike is substantially larger than zero (0.53) in this study, we observe very high empirical bias in the SP model-based estimates. However, our methods provided negligible empirical biases even in the presence of a close spike. We also presented the estimated number of distant spikes in each of the simulation studies in Table B.1. Note that in some cases our algorithm over-estimates the number of distant spikes. However, as the over-estimation is only finite, the estimates of the distant spikes still remain consistent.

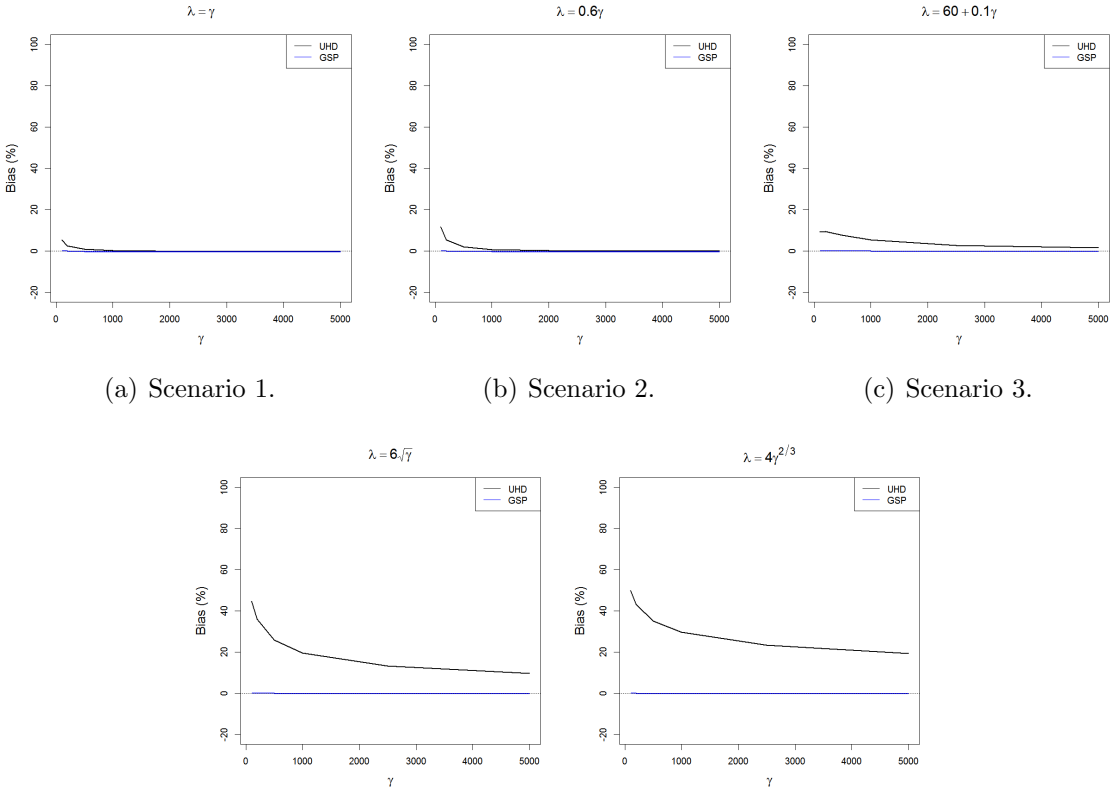
2.6.2 Simulation Studies: Compare GSP and Ultra High-Dimensional (UHD) Regime-Based Methods

In Section 2.4.6 we compared the asymptotic results under the UHD regime and the results based on the high-dimensional GSP model when p is greatly larger than n, but p/n is large but finite. We theoretically established that the results from the two regimes become almost identical when $p/n = \gamma$ is sufficiently large. However,

given large but finite γ in the data, the difference can be substantial when the spike is smaller compared to γ. In this section, we will assess that result by numerically comparing the GSP and UHD-based estimates for different values of γ. We considered five different scenarios where the largest population eigenvalue $\lambda = \gamma, 0.6\gamma, 60 + 0.1\gamma, 6\sqrt{\gamma}, 4\gamma^{2/3}$. For the first three scenarios, under the UHD regime, λ/γ can be assumed to be bounded away from zero, and for the last two, $\lambda/\gamma \to 0$ as $\gamma \to \infty$. To compare the performances as γ increases, we selected six different values for $\gamma = 100, 200, 500, 1000, 2500, 5000$. For each combination of γ and λ, we simulated 200 datasets, each with $n = 200$ samples from a population with only one spike λ, and the non-spikes generated from the AR(1) covariance structure with $(\sigma^2, \rho) = (1, 0.9)$.

First, we compare the convergence results of the largest sample eigenvalue d from Theorem 2.1 and (2.5). For this purpose, we assume the population eigenvalues and the rate of increment of λ are known, and we compare the relative errors $\epsilon_{GSP} = (d - \psi(\lambda))/d$ and $\epsilon_{UHD} = (d - \lambda - \gamma)/d$ or $(d - \gamma)/d$ depending on whether λ/γ is assumed to be bounded away from zero or not. Figure 2.3 shows that for all combinations of (γ, λ), the GSP-based convergence result (Theorem 2.1) has very negligible relative errors. On the other hand, the UHD-based convergence result (2.5) has substantially large relative errors even for γ as large as 5000 in scenarios 3, 4 and 5. For scenarios 1 and 2, since λ increases at a faster rate with γ than in other scenarios, the relative errors based on the two results converge much faster. However, for relatively smaller values of γ (100, 200, 500), the differences are substantial even though γ is large compared to $n = 200$. This suggests that we need γ to be large in an absolute sense, and not only in a relative sense compared to n in order to assume $\gamma \to \infty$ and apply UHD-based results.

Next, we compare the estimates of the spike λ using GSP-based and UHD-based methods assuming the population eigenvalues and the rate of increment of the population spike λ to be unknown. Among the GSP-based methods we only used the



(a) Scenario 1. (b) Scenario 2. (c) Scenario 3.

(d) Scenario 4. (e) Scenario 5.

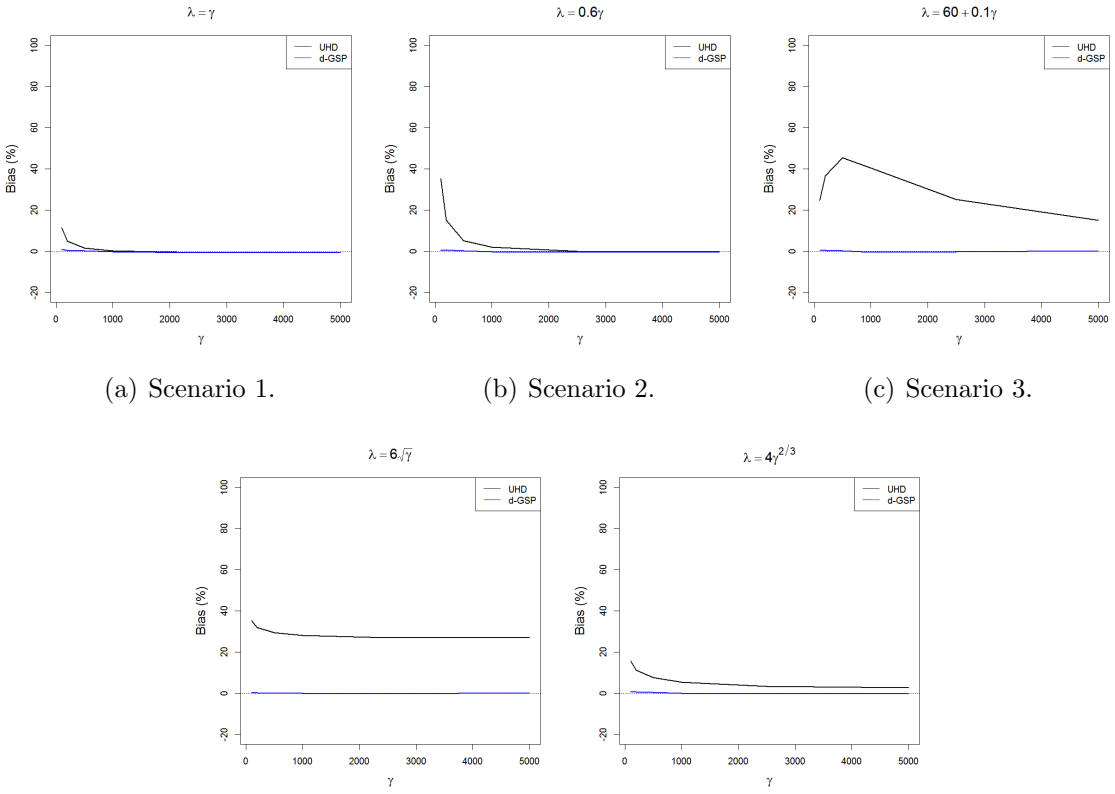
Figure 2.3: Comparison of the relative errors (%) in the convergence results of the largest sample eigenvalue derived under the GSP and UHD assumptions. The population eigenvalues and the rate of increment of the largest population eigenvalue are assumed to be known.

d-GSP method for this purpose due to the computational burden associated with applying the λ-GSP method on such a large number of simulated datasets. One thing to note here is that the UHD results do not provide any consistent estimators for λ, as it is assumed to be divergent when $\lambda \geq O(\gamma)$, and the asymptotic properties of the sample eigenvalues do not depend on λ when $\lambda/\gamma \to 0$. Thus, in order to compare these methods, we estimate λ by $\hat{\lambda} = d - \gamma$ when considering the UHD regime. From Figure 2.4 we can see that our proposed d-GSP method provides almost negligible biases for all combinations of (γ, λ), whereas the UHD-based estimates have substantial biases even for γ as large as 5000 in scenarios 3, 4 and 5. For scenario 1 and 2, both methods provide almost unbiased estimates when $\gamma \geq 1000$ and $\gamma \geq 2000$ respectively. Further, we compare the estimated shrinkage factors based on these two methods in Figure B.1. They also show very similar patterns as the estimated spikes.

2.6.3 Application on Hapmap III Data

For this demonstration we used genetic data from the Hapmap Phase III project (`http://hapmap.ncbi.nlm.nih.gov/`). Our sample consisted of unrelated individuals sampled from two different populations: a) CEU and b) TSI. We only included genomic markers that are on chromosome 1-22, have less than 5% missing values, and those with minor allele frequency more than 0.05. We also excluded 2 samples (both from CEU) with outlier PC scores (more than six standard deviations away from the mean PC score corresponding to at least one distant spike). We then mean-centered and variance-standardized the data for each marker. The final sample consisted of 198 individuals (110 from CEU and 88 from TSI). Total number of markers selected across chromosome 1-22 was 1389511.

To evaluate the performance of the proposed methods with different p, we performed PCA on each chromosome separately. The number of markers varied from



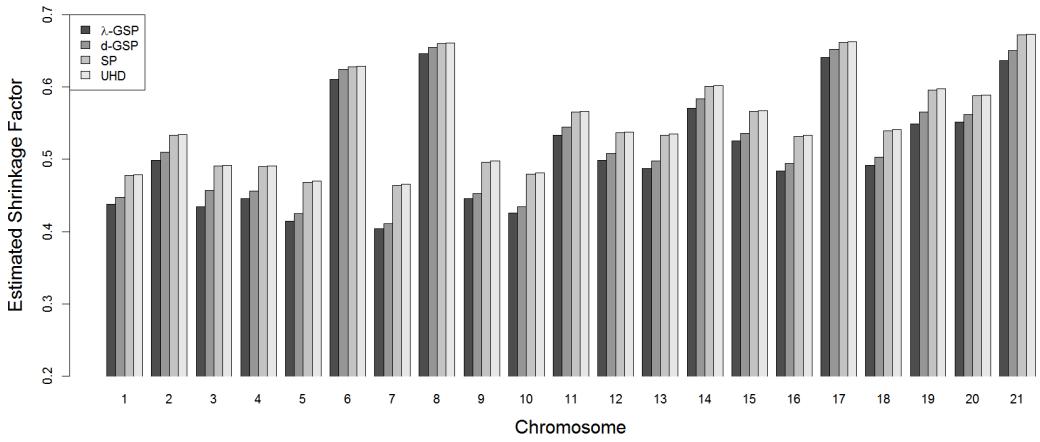
(a) Scenario 1. (b) Scenario 2. (c) Scenario 3.

(d) Scenario 4. (e) Scenario 5.

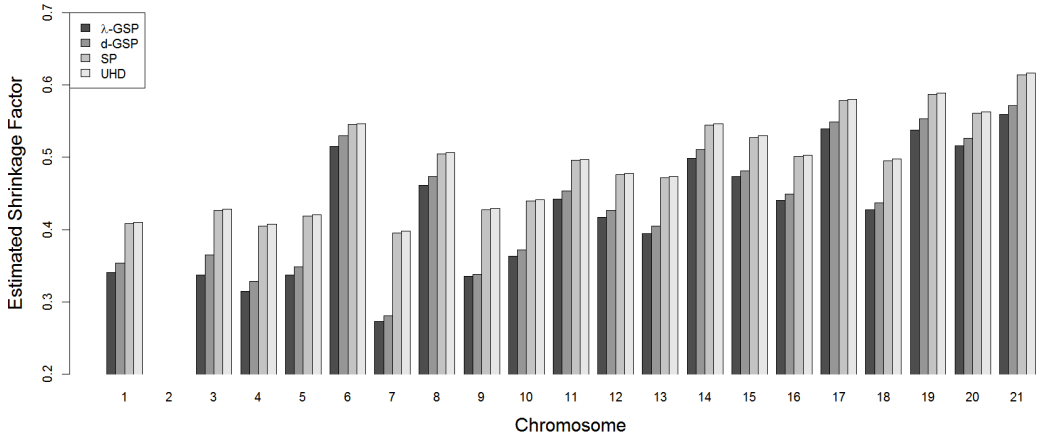
Figure 2.4: Empirical biases (%) in estimating the largest population eigenvalue for GSP-based and UHD-based methods. The population eigenvalues and the rate of increment of the largest population eigenvalue are assumed to be unknown.

19331 (chromosome 21) to 116582 (chromosome 2). The distribution of the number of markers across different chromosomes are presented in Figure B.3. We first estimated the number of distant spikes using the algorithm described in Section 2.5.3. We found no distant spike in chromosome 22 and only one distant spike in chromosome 2. Then we applied our GSP-based methods, the existing SP-based method (*Lee et al.*, 2010) and the UHD-based method (*Lee et al.*, 2014b) to estimate the asymptotic shrinkage factors corresponding to the distant spikes. Figures 2.5(a), 2.5(b) compares the estimated asymptotic shrinkage factors for the first two PCs across different chromosomes. The plots show that for all the chromosomes, λ-GSP and d-GSP methods provided almost equal estimates while the SP and UHD estimates are larger than both the GSP estimates. This suggests that the SP method would over-estimate the shrinkage factors when the population eigenvalues deviate from the assumption that the non-spiked eigenvalues are the same. Moreover, the UHD method over-estimated the shrinkage factors even for p/n nearly as large as 600 (chromosome 2).

To investigate whether the proposed shrinkage-bias adjustment can improve the prediction accuracy, we performed a leave-one-out cross-validation. In each iteration we removed one individual (test sample) and performed PCA on the remaining individuals (training samples) to predict the PC score of the test sample. For each predicted PC score, we adjusted the shrinkage-bias using the GSP-based, SP-based and UHD-based shrinkage factor estimates. One important issue with this cross-validation is that the exclusion of one individual can substantially change the PC-coordinates, in which the PC score plots from the training sample-based and complete sample-based PCA can be substantially different. In order to circumvent this problem, in each iteration we first rescaled the PC scores based on their corresponding sample eigenvalues to make the PCs comparable. In addition, we obtained the mean squared difference of the training sample PC1-2 scores with and without the exclusion of the test sample (for chromosome 2, only PC1 is used), and excluded the test sample from



(a) Shrinkage factors for PC1.



(b) Shrinkage factors for PC2.

Figure 2.5: Comparison of the estimated shrinkage factors using different methods on the Hapmap data.

the prediction error estimation if the mean squared difference was above a threshold ϵ. We used four different values $0.5, 1, 5$ and 10 for the threshold parameter ϵ, and for each value of ϵ we calculated the MSE of the unadjusted and adjusted predicted PC scores of the test samples. The sample sizes of the test samples that were finally included in the prediction error estimation for different values of ϵ are shown in Figure B.2. Figure 2.6 shows the estimated MSEs for $\epsilon = 1$. It is clear that both the λ-GSP and d-GSP methods have much smaller MSEs than the SP method. The UHD-based method had almost identical MSEs as the SP-based method for all the chromosomes, hence we omitted the UHD-based results in this plot. As expected, the unadjusted predicted PC scores have substantially larger MSE than all the proposed adjustments. The plots are very similar for the other values of ϵ, and they can be found in Figure B.4.

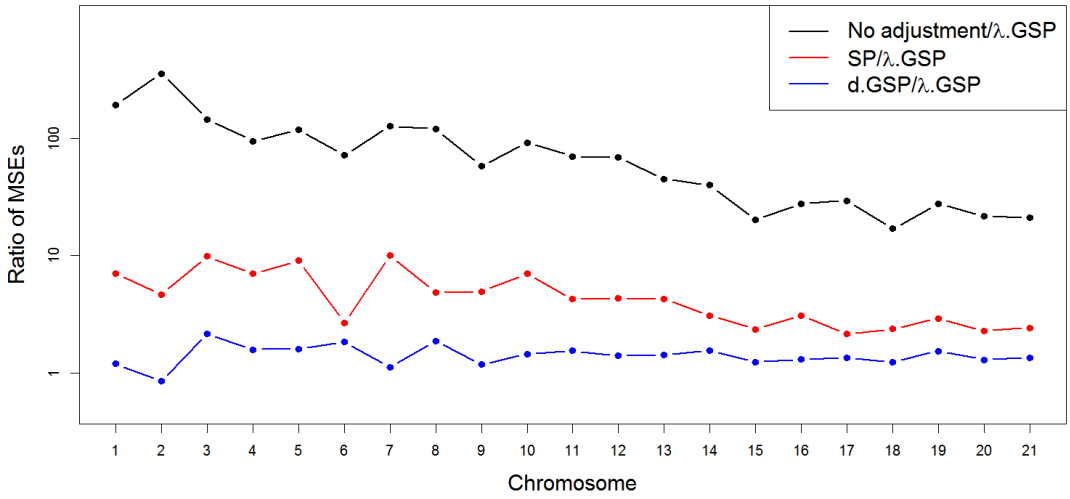


Figure 2.6: Comparison of the MSE of the unadjusted, d-GSP-adjusted, and SP-adjusted PC scores, with the λ-GSP-adjusted PC scores using $\epsilon = 1$. The ratios of the MSEs are presented for chromosome 1-21. The Y-Axis is presented in a logarithmic scale.

Figure 2.7 illustrates the shrinkage-bias adjustment for the PC1 and PC2 scores of an individual based on the markers on chromosome 7. The plot clearly shows

that the bias-adjusted PC score based on the SP model is still biased towards zero, whereas the bias-adjusted PC score based on the GSP model is very close to the original sample PC score. We only showed the d-GSP adjusted score in the plot as the d-GSP and λ-GSP adjusted scores were almost equal.

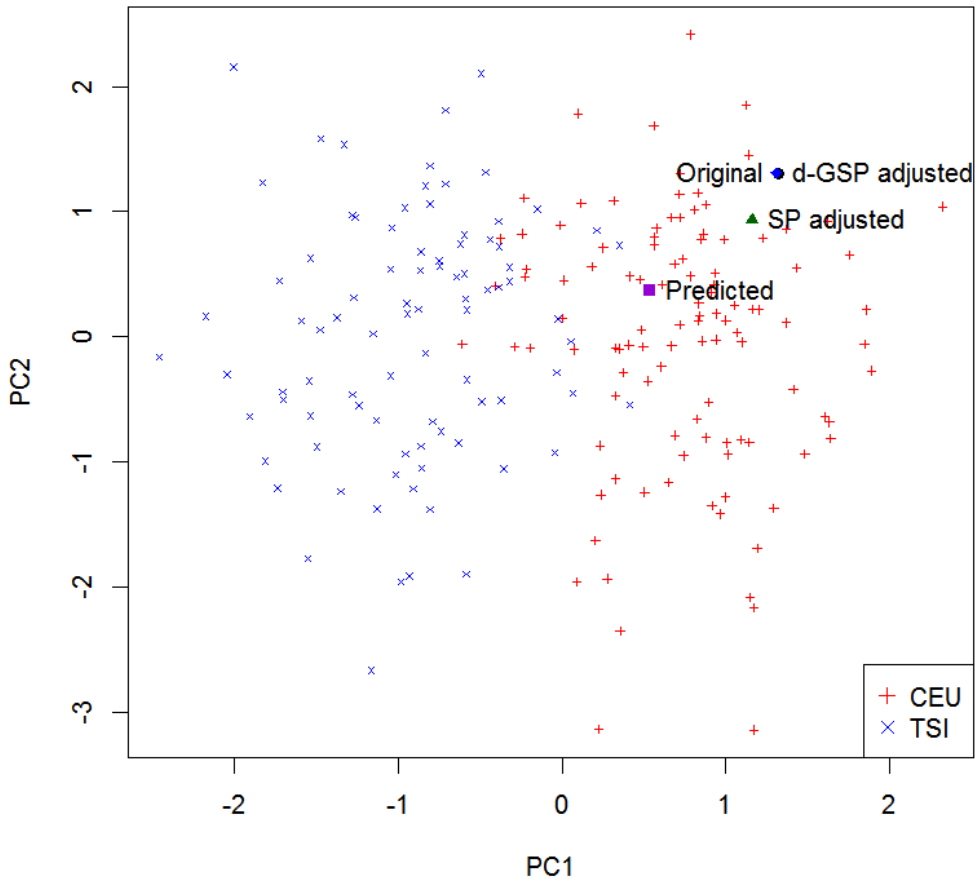


Figure 2.7: PC1 vs PC2 plot of the Hapmap III CEU and TSI samples based on chromosome 7. The predicted PC scores for the illustrative individual, and its bias-adjusted PC scores are also presented. Since the d-GSP and the λ-GSP adjusted scores are nearly the same, the λ-GSP adjusted scores are not presented.

2.7 Discussion

We investigated the asymptotic properties of PCA under the Generalized Spiked Population model and derived estimators of the population eigenvalues, the angles between the sample and population eigenvectors, and the correlation coefficients between the sample and population PC scores. We also proposed methods to adjust the shrinkage bias in the predicted PC scores. Further, theoretically and using simulation studies, we compared our results with the results developed under the ultra high-dimensional regime (*Lee et al.*, 2014b), and showed that our methods provide more accurate estimates when $p/n \to \gamma$ is asymptotically finite but large compared to n in the given data, and the spike of interest is small compared to γ. When the spike is large compared to γ, both methods provide nearly identical estimates. Since the proposed methods do not require the equality of the non-spiked eigenvalues, they can be widely used in high-dimensional biomedical data analysis. We implemented all our algorithms in the R package **hdpca**.

We note that *Mestre* (2008a,b) proposed an asymptotic setting similar to the generalized spiked population model but with a different assumption on the number of spikes in which the number of spikes increases with the dimension. Under this assumption, he provided asymptotic properties of sample eigenvalues and eigenvectors. However, in many biomedical data, the number of spikes is usually finite as the spikes represent the difference between finitely many underlying subpopulations. Therefore we believe that the generalized spiked population model is more appropriate in such cases.

In some special cases, even though the features exhibit strong local correlation, one can use the spiked population model based methods after some suitable data manipulation. In genome-wide association studies, SNP pruning (*Anderson et al.*, 2010) can be used to remove locally correlated SNPs to satisfy the spiked population model. For example, *Lee et al.* (2010) reported good performance of the spiked

population model-based methods with the SNP-pruned Hapmap III dataset. This approach, however, can lead to a considerable loss of information; the SNP-pruning in Hapmap III data removed nearly 90% of the SNPs. Since the proposed approach does not require this additional step, it can use most of the information present in the data.

CHAPTER III

Two-Stage PLS Method to Address the Over-Fitting Problem in Partial Least Squares Regression on High-Dimensional Predictors

3.1 Introduction

Partial least squares (PLS) is one of the most widely used multivariate statistical methods for dimension reduction in regression models. Originally developed by *Wold* (1966, 1982) to address the econometric path modeling problems in the social sciences literature, PLS is now widely popular in the field of spectroscopy and chemometrics (*Wold et al.*, 1984, 2001; *de Jong*, 1993; *Geladi and Kowalski*, 1986), and also has been applied in many biomedical fields including genomics, metabolomics, neurology etc. (*Boulesteix and Strimmer*, 2007; *Man*, 2004; *Huang et al.*, 2005; *Clementi et al.*, 1997; *McIntosh et al.*, 1996) due to its attractive ability of handling large number of predictors and modeling multiple outcomes simultaneously. It is especially useful in problems where the predictors are highly correlated among themselves.

PLS is a closely related technique to principal component analysis (PCA) (*Helland*, 1990; *Stoica and Söderström*, 1998). In fact, PLS combines PCA and multiple linear regression (MLR) methods to simultaneously achieve dimension reduction and model fitting. Because of this dimension reduction feature of PLS, it is an attractive

tool to be applied in data with high-dimensional predictors, where the number of predictors is larger than the sample size. The latent structure model employed by PLS can also provide a framework for identifying and adjusting for unknown confounders in regression models. For example, *Epstein et al.* (2007) proposed a method using PLS to control for population stratification in genetic association studies.

In models with high-dimensional predictors, however, PLS can suffer from an over-fitting problem, where the fitted and observed outcomes are almost identical, even when the outcomes are completely independent of the predictors. To illustrate this phenomenon, we simulated $n = 500$ subjects with outcomes $Y_i \sim N(0, 1)$ for i^{th} subject, and independently simulated $p = 10000$ covariates $X_{ij} \sim N(0, 1), j = 1, \ldots, p$ for each subject. Then we used PLS with one component to fit this regression model. The resulting squared correlation (R^2) between the fitted outcomes ($\hat{Y}$) and observed outcomes (Y) was 0.951 (Figure 3.1), even though Y and X were generated independently. This can result in falsely inferring that a large proportion of variability in the outcomes can be explained by the predictors, or falsely attributing a higher effect of the predictors on the outcomes, when the true effect is substantially lower. The over-fitting problem has also been observed and discussed by *Brereton and Lloyd* (2014); *Gromski et al.* (2015) with respect to fitting classification models using the PLS method. *Lee et al.* (2008); *He et al.* (2017) also raised concerns regarding the use of PLS for confounder adjustment, for being prone to over-fitting the model, which can further result in spurious confounders and loss of power.

Even though the over-fitting problem of PLS has previously been noticed by others, to the best of our knowledge, no method has been developed to properly address this problem. The sparse PLS (SPLS) method (*Chun and Keleş*, 2010) was previously proposed for high-dimensional data when the predictors are sparse, i.e, only a small number of predictors are relevant. The performance of this method, in terms of addressing the over-fitting problem, also needs to be evaluated under different sparsity

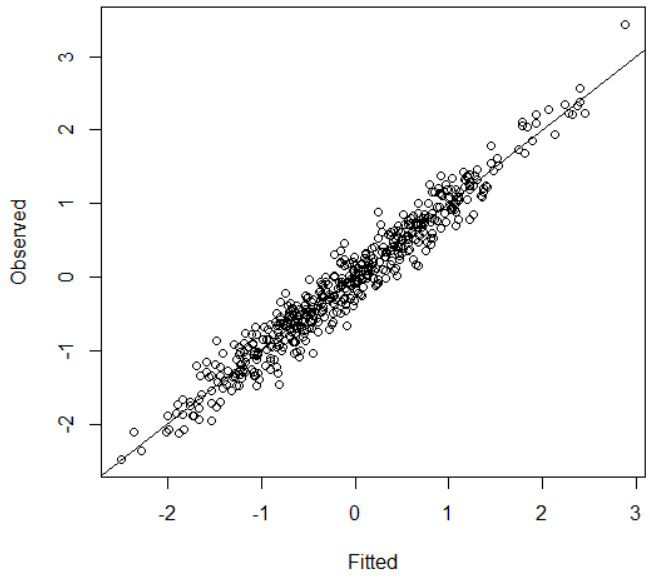


Figure 3.1: Fitted vs observed outcomes for PLS regression with independently generated outcomes and predictors.

levels.

In this chapter, we first investigate the reasons behind the over-fitting problem, and propose a two-stage PLS (TPLS) method to address the problem using the recent developments in the random matrix literature for high-dimensional PCA (*Lee et al.* (2010), and Chapter II). Since, PCA and PLS are closely related methods, the shrinkage phenomenon discussed in *Lee et al.* (2010) can also affect the prediction performances of our method. We further investigate the effect of shrinkage on PLS, and incorporate the shrinkage-bias adjustment techniques presented in *Lee et al.* (2010) and in Chapter II into our method. We evaluate and compare our proposed method with the traditional PLS and the SPLS methods under various simulated scenarios, as well as apply the proposed method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data to demonstrate its performance.

3.2 Over-Fitting in High-Dimensional Partial Least Squares

We consider the linear regression model of $Y = XB + G$ with n independent samples, p predictors, and q outcomes, where $Y_{n\times q} = [y_1 \quad y_2 \quad \dots \quad y_q]$ is the matrix of outcomes, $X_{n\times p} = [x_1 \quad x_2 \quad \dots \quad x_p]$ is the matrix of predictors, $B_{p\times q}$ is the matrix of coefficients, and $G_{n\times q}$ is the random error matrix. The PLS method estimates the parameters of the linear regression model by assuming the following latent decompositions,

$$X = TP^T + E$$
$$Y = TQ^T + F,$$

where T is an $n \times k$ matrix of k latent scores, $rank(T) = k$, and $P_{p\times k}$ and $Q_{q\times k}$ are the corresponding loading matrices, respectively. $E_{n\times p}$ and $F_{n\times q}$ are the random error matrices. The rows of E have mean zero and variance $\sigma_x^2 I_p$, and the rows of F have mean zero and variance $\sigma_y^2 I_q$, where I_p and I_q are the identity matrices of order p and q, respectively. We assume q and k to be finite, and X to be high-dimensional, $p \to \infty, n \to \infty, p/n \to \gamma < \infty$, $\gamma > 1$.

The PLS method estimates the latent scores using $\hat{T} = XW$, where the columns of the $W_{p\times k} = [w_1 \quad w_2 \quad \dots \quad w_k]$ are found by solving the following optimization problem,

$$\begin{aligned} w_i = \arg\max_w n^{-2} w^T X^T Y Y^T X w \quad \text{s.t.} \quad & w^T w = 1, \\ & w^T S_{xx} w_j = 0, \forall j = 1, \dots, i-1, \end{aligned} \tag{3.1}$$

where $S_{xx} = X^T X/n$. Here, we assumed without loss of generality, that the columns of X and Y are mean-centered. Then, the columns of Y are regressed on $\hat{T}$ using ordinary least squares (OLS) method to obtain $\hat{Q}$, the estimate for Q. Finally, the linear regression coefficients are estimated by $\hat{B} = W\hat{Q}^T$.

Notice that, TP^T is the structural part, or signal part of X which affects the structural part TQ^T of the outcome under the model assumptions. Ideally, in each successive optimization, we want to find the direction w which maximizes the sample covariances between TP^Tw and columns of TQ^T, i.e, to solve the following optimization problem,

$$w_i = \arg\max_w n^{-2} w^T P T^T T Q^T Q T^T T P^T w \quad \text{s.t.} \quad w^T w = 1, \\ w^T S_{xx} w_j = 0, \forall j = 1, \ldots, i-1. \tag{3.2}$$

We first show that under the low-dimensional setting, where p is finite, (3.1) and (3.2) are asymptotically equivalent optimization problems. We can decompose the objective function of (3.1) by,

$$n^{-2} w^T X^T Y Y^T X w = n^{-2} w^T P T^T T Q^T Q T^T T P^T w + n^{-2} w^T P T^T F F^T T P^T w \\ + n^{-2} w^T E^T Y Y^T E w + 2 n^{-2} w^T E^T Y Y^T T P^T w \tag{3.3}$$

Now, consider the term,

$$n^{-2} w^T E^T Y Y^T E w = \sum_{j=1}^{q} Cov^2(y_j, Ew) = \sum_{j=1}^{q} \left(\sum_{i=1}^{p} Cov\left(y_j, E_i w^{(i)}\right) \right)^2,$$

where E_i is the i^{th} column of E, and $w^{(i)}$ is the i^{th} element of w. For any non-random w such that $w^T w = 1$, since E and Y are independent, each $Cov\left(y_j, E_i w^{(i)}\right) \xrightarrow{a.s.} 0$. When p is finite, this implies $n^{-2} w^T E^T Y Y^T E w \xrightarrow{a.s.} 0$. The same observation can also be made for the term $2 n^{-2} w^T E^T Y Y^T T P^T w$. Moreover, since F only has q many columns, regardless of p, the term $n^{-2} w^T P T^T F F^T T P^T w$ goes to zero almost surely. Therefore, under the low-dimensional setting, solving (3.1) provides asymptotically

identical solutions as (3.2).

However, in the high-dimensional setting, the last two terms of (3.3) are not guaranteed to converge to zero as they include p many sums. Moreover, since the column-space of X, which is a full row rank matrix of order $n \times p$, will always contain the columns of Y, the canonical correlation between X and Y will always be unity. Thus, when maximizing $n^{-2}w^T X^T Y Y^T X w$, the PLS method may not maximize $n^{-2}w^T P T^T T Q^T Q T^T T P^T w$, and instead can over-fit the model by increasing the values of the last two terms of the right-hand side (RHS) of (3.3). For example, suppose all elements of Q are zeros, which means X and Y are completely unrelated. Then, because the columns of Y belong to the column space of X (full row rank matrix), solving (3.1) will result in selecting w-s such that each Xw belongs to the column space of Y, and after selecting q components, $X\,[w_1 \quad \dots \quad w_q]$ will span the same column space as Y. Obviously, when $q = 1$, Xw_1 will be a scalar multiple of Y. Thus, when fitting Y on $\hat{T} = XW$, the fitted outcomes will be identical to the observed outcome Y, and the coefficient of determination (R^2) in the model will be unity, which is an obvious case of over-fitting.

3.3 Two-stage PLS (TPLS) Method

First, we note that the over-fitting issue is solved if w is restricted to any finite dimensional subspace $\mathscr{C}(S)$ (column-space of S), where S is a $p \times m$ matrix and m is finite. For any $w_{p\times 1} \in \mathscr{C}(S)$, let $w = S\tilde{\gamma}$, and $ES = \tilde{E}$. Then, $n^{-1}Y^T E w = n^{-1}Y^T \tilde{E}\tilde{\gamma} = \left(\sum_{i=1}^m Cov(y_1, \tilde{E}_i\tilde{\gamma}_i), \dots, \sum_{i=1}^m Cov(y_q, \tilde{E}_i\tilde{\gamma}_i)\right)$, where $\tilde{E}_i$ is the i^{th} column of $\tilde{E}$, and $\tilde{\gamma}_i$ is the i^{th} element of $\tilde{\gamma}$. As Y and $\tilde{E}$ are independent, each sample covariance $Cov(y_j, \tilde{E}_i\tilde{\gamma}_i) \xrightarrow{a.s.} 0$. When m is finite, this implies $n^{-1}Y^T E w \xrightarrow{a.s.} 0$ and the last two terms on the RHS of (3.3) converges almost surely to zero. On the other hand, if m is not finite, then this convergence is not guaranteed, resulting in possible over-fitting of the model. The SPLS method (*Chun and*

Keleş, 2010) also addresses the over-fitting issue when sparsity assumptions (as noted in *Chun and Keleş* (2010)) are valid, as it also imposes a subspace constraint where the subspace is spanned by finitely many euclidean basis vectors. However, when the sparsity assumption is not applicable, the euclidean subspace constraint may not be optimum, or it may result in over-fitting the model by selecting a large number of variables.

In order to find the optimum finite-dimensional subspace constraint for solving the optimization problem (3.2), we introduce the following theorem, that discusses the nature of the solutions to (3.2),

Theorem 3.1. *Let* $\tilde{U} = [u_1 \quad u_2 \quad \dots \quad u_k]$*, where* u_i *is the eigenvector corresponding to* λ_i*, the* i^{th} *largest eigenvalue of* $\Sigma_p = n^{-1}PT^TTP^T$*, and* w_i *is the* i^{th} *successive solution to the optimization problem* (3.2)*. Then,* $w_i \in \mathscr{C}\left(\tilde{U}\right)$.

The proof can be found in Appendix C. Theorem 3.1 shows that any solution to (3.2) will belong to the subspace $\mathscr{C}\left(\tilde{U}\right)$. Now, $\tilde{U}$ comprises of the first k eigenvectors of Σ_p, which are the same as the first k eigenvectors of $\Gamma_p = \Sigma_p + \sigma_x^2 I_p$, the population covariance matrix of X. Notice that, Γ_p follows the spiked population model as described in *Johnstone* (2001) with first k eigenvalues as spikes, assuming the largest eigenvalue of Σ_p to be bounded. Under the spiked population model in the high-dimensional setting, even though we cannot consistently estimate the population eigenvectors, we can consistently estimate the angles between the sample and population eigenvectors using theoretical results derived in *Lee et al.* (2010), and also in Chapter II as a special case of the generalized spiked population model.

For the convenient use of notations, we denote the eigenvalues of Γ_p as $\theta_1, \dots, \theta_p$ in decreasing order of magnitude; $\theta_i = \lambda_i + \sigma_x^2$ for $i = 1, \dots, k$, and $\theta_i = \sigma_x^2$ for $i = k+1, \dots, p$. Let $S_{xx}/n = VDV^T$ be the eigendecomposition of the sample covariance matrix, where D is diagonal with eigenvalues $d_1 \geq d_2 \geq \dots \geq d_p$, and $V = [v_1 \quad v_2 \quad \dots \quad v_p]$ comprises of the sample eigenvectors. Further, we assume the

following,

Assumption 3.1. *As we are interested only in the first k eigenvectors of Γ_p, and the eigenvectors do not change if we scale all the sample and population eigenvalues by any non-zero scalar quantity, without loss of generality we assume $\sigma_x^2 = 1$.*

Assumption 3.2. *The multiplicity of $\theta_1, \ldots, \theta_{k^*}$ are all unity, where k^* is the number of eigenvalues of Γ_p larger than $1 + \sqrt{\gamma}$.*

Then, based on Theorem 2.2 in Chapter II, it can be shown that when $\theta_i > 1+\sqrt{\gamma}$,

$$\left| \langle v_i, u_i \rangle^2 - \frac{\theta_i \psi'(\theta_i)}{\psi(\theta_i)} \right| \xrightarrow{p} 0, \qquad \langle v_i, u_j \rangle^2 \xrightarrow{p} 0, \quad j \neq i. \tag{3.4}$$

where $\psi(\theta) = \theta\,(1 + \gamma/(\theta - 1))$, and ψ' is the derivative of ψ. Notice that, as per the terminology defined in Chapter II, the spikes $(\theta_1, \ldots, \theta_{k^*})$ larger than $1 + \sqrt{\gamma}$ are called the distant spikes, the spikes $(\theta_{k^*+1}, \ldots, \theta_k)$ smaller than $1 + \sqrt{\gamma}$ are called the close spikes, and the rest of the population eigenvalues $(\theta_{k+1}, \ldots, \theta_p)$ are called non-spikes. The following theorem provides the convergence of the angles between the population eigenvectors corresponding to the spikes and the sample eigenvectors corresponding to the close spikes and non-spikes.

Theorem 3.2. *Let v_i be a sample eigenvector such that the corresponding population eigenvalue, $\theta_i \leq 1 + \sqrt{\gamma}$, and the sample eigenvalue $d_i > 0$. If $\gamma > 1$, then, for all $j = 1, \ldots, k$,*

$$\langle v_i, u_j \rangle \xrightarrow{p} 0.$$

The proof can be found in Appendix C.

Theorem 3.2 shows that the low-rank sample eigenvectors v_is such that $i > k^*$ and $d_i > 0$ are all asymptotically orthogonal to $\mathscr{C}\left(\tilde{U}\right)$. Moreover, the eigenvectors corresponding to the sample eigenvalues $d_i = 0$ does not provide any information as the sample predictors do not have any variability on those directions. Therefore, we

propose to remove all low-rank eigenvectors from the universal p-dimensional space, and restrict w to the remaining subspace, which is spanned by only the first k^* sample eigenvectors of S_{xx}/n. Because $\tilde{V} = [v_1 \quad \dots \quad v_{k^*}]$ only has finitely many columns, our approach will solve the over-fitting problem. This selection of sample eigenvectors is the best possible, in the sense that, adding finitely many low-rank eigenvectors to the basis will not improve the estimation, as the additional eigenvectors provide no information on $\mathscr{C}\left(\tilde{U}\right)$. On the other hand, including infinitely many of them in the basis can result in over-fitting of the model. This implies, in the two-stage PLS (TPLS) method, we can first calculate the first k^* sample principal component (PC) scores $X^* = X\tilde{V}$ (the first stage). The number of distant spikes k^* can be estimated using the algorithm described in *Lee et al.* (2010) (Section 2.4), or a more general algorithm described in Chapter II (Section 2.5.3). Then, we solve the optimization problem,

$$\gamma_i^* = \arg\max_{\gamma_i^*} n^{-2}\gamma_i^{*T}X^{*T}YY^TX^*\gamma_i^* \quad \text{s.t.} \quad \gamma_i^{*T}\gamma_i^* = 1,$$
$$\gamma_i^{*T}\tilde{V}^TS_{xx}\tilde{V}\gamma_j^* = 0, \forall j = 1, \dots, i-1,$$

which is equivalent to performing the PLS regression of Y on X^* (the second stage).

3.4 Consistent Estimation of the Variability in Y Explained by X

As shown in our simulated example (Figure 3.1), over-fitting a PLS model can result in falsely inferring that the predictors (X) explain almost all of the variability in the outcomes (Y), even when they are independent. To properly understand how much effect the predictors truly have on the outcomes, it is thus important to consistently estimate of the maximum proportion of variability (R^2) in Y, that can be explained by X without over-fitting the model. In particular, we need to estimate the

expected maximum variability in Y (sum of the column-wise variances of the fitted $\hat{Y}$) that can be explained by X, in the ideal situation where the selection of ws are constrained to the optimum subspace $\mathscr{C}\left(\tilde{V}\right)$.

However, $\tilde{U}$ is not known, and in our TPLS method, we are restricting our selection of w-s to the subspace $\mathscr{C}\left(\tilde{V}\right)$ instead of $\mathscr{C}\left(\tilde{U}\right)$. As $\mathscr{C}\left(\tilde{V}\right)$ is not consistent to $\mathscr{C}\left(\tilde{U}\right)$ in the high-dimensional setting, the solution for w will be sub-optimal for (3.2), which can lead to a loss of variability in $\hat{Y}$, compared to the optimal solution. Even though we cannot obtain the optimal solution to (3.2), in this section, we will derive an asymptotically unbiased estimate for the expected maximum variability in Y that can be explained by X, using the theoretical results derived in *Lee et al.* (2010) and Chapter II.

Notice that, restricting the PLS directions w-s to the subspace $\mathscr{C}\left(\tilde{U}\right)$ is equivalent to performing a PLS of Y on $\tilde{X} = X\tilde{U}$, and the variability in $\hat{Y}$ is maximum when all k components are selected, which is equivalent to performing OLS regressions of the columns of Y individually on $\tilde{X}$. Similar observation also holds for the PLS regression of Y on X^* with k^* components. Without loss of generality, we assume Y only has one column ($q = 1$). We are interested in estimating the expected explained variance of Y in the following true underlying model,

$$Y = \tilde{X}\tilde{\eta} + \tilde{\epsilon}, \tag{3.5}$$

using the parameter estimates from the misspecified model,

$$Y = X^*\eta^* + \epsilon^*. \tag{3.6}$$

$\tilde{\eta}, \eta^*$ are the coefficients corresponding to $\tilde{X}$ and X^*, and $\tilde{\epsilon}, \epsilon^*$ are the error terms in the two models, $\tilde{\epsilon} \sim N(0, \sigma_y^2)$. If Y has more than one columns, then we can estimate the explained variance for each column individually and sum them up.

We first assume that all the spikes of Γ_p are distant spikes (which implies $k^* = k$), and k^* is correctly estimated. Let the Assumptions 3.1 and 3.2 hold. Then, according to *Lee et al.* (2010), if $\theta_i > 1 + \sqrt{\gamma}$,

$$|d_i - \psi(\theta_i)| \xrightarrow{p} 0. \tag{3.7}$$

Without loss of generality, we assume $\langle v_i, u_i \rangle$ is non-negative for all $i = 1, \ldots, k$. Then, from (3.4) and (3.7),

$$\begin{aligned}
\tilde{X}^T \tilde{X}/n &= \tilde{U}^T \left(S_{xx}/n\right) \tilde{U} \xrightarrow{p} \Theta_k \\
X^{*T} X^*/n &= \tilde{V}^T \left(S_{xx}/n\right) \tilde{V} = \tilde{V}^T V D V^T \tilde{V} = D_k \xrightarrow{p} \Psi_k, \quad \text{and} \\
\tilde{X}^T X^*/n &= \tilde{U}^T \left(S_{xx}/n\right) \tilde{V} = \tilde{U}^T V D V^T \tilde{V} \xrightarrow{p} \left[\Theta_k \Psi_k' \Psi_k\right]^{1/2},
\end{aligned}$$

where Θ_k is a diagonal matrix with diagonal elements $\theta_1, \ldots, \theta_k$, D_k is a diagonal matrix with diagonal elements $d_1, \ldots, d_k$, Ψ_k is a diagonal matrix with diagonal elements $\psi(\theta_1), \ldots, \psi(\theta_k)$, and Ψ_k' is a diagonal matrix with diagonal elements $\psi'(\theta_1), \ldots, \psi'(\theta_k)$. The first convergence holds because $S_{xx}/n \xrightarrow{p} \Sigma_p$ element-wise.

Let $\tilde{P}_X = \tilde{X} \left(\tilde{X}^T \tilde{X}\right) \tilde{X}^T$ and $P_X^* = X^* \left(X^{*T} X^*\right) X^{*T}$. Then the fitted outcomes based on models (3.5) and (3.6) are given by $\tilde{Y} = \tilde{P}_X Y$ and $Y^* = P_X^* Y$, respectively. The corresponding explained variances are given by $\tilde{H}_E = n^{-1} \tilde{Y}^T \left(I_n - J_n/n\right) \tilde{Y}$, and $H_E^* = n^{-1} Y^{*T} \left(I_n - J_n/n\right) Y^*$, where J_n is the $n \times n$ matrix with all elements equal

to unity. Then, under the true model (3.5),

$$\begin{aligned}
E\left(\tilde{H}_E\right) &= E\left(n^{-1}\tilde{Y}^T\left(I_n - J_n/n\right)\tilde{Y}\right) \\
&= E\left(n^{-1}Y^T\tilde{P}_X\left(I_n - J_n/n\right)\tilde{P}_X Y\right) \\
&= E\left(n^{-1}Y^T\tilde{P}_X Y\right) \quad \text{as } \tilde{X}^T J_n = 0 \\
&= n^{-1}tr\left(\tilde{P}_X\sigma_y^2\right) + n^{-1}\tilde{\eta}^T\tilde{X}^T\tilde{P}_X\tilde{X}\tilde{\eta} \\
&= (k/n)\,\sigma_y^2 + n^{-1}\tilde{\eta}^T\tilde{X}^T\tilde{X}\tilde{\eta} \\
&\xrightarrow{p} \tilde{\eta}^T\Theta_k\tilde{\eta} \quad \text{as } n \to \infty \\
E\left(H_E^*\right) &= E\left(n^{-1}Y^{*T}\left(I_n - J_n/n\right)Y^*\right) \\
&= E\left(n^{-1}Y^T P_X^*\left(I_n - J_n/n\right)P_X^* Y\right) \\
&= E\left(n^{-1}Y^T P_X^* Y\right) \quad \text{as } X^{*T} J_n = 0 \\
&= n^{-1}tr\left(P_X^*\sigma_y^2\right) + n^{-1}\tilde{\eta}^T\tilde{X}^T P_X^*\tilde{X}\tilde{\eta} \\
&\xrightarrow{p} \tilde{\eta}^T\Theta_k\Psi_k'\tilde{\eta} \quad \text{as } n \to \infty
\end{aligned}$$

This clearly shows that $E\left(\tilde{H}_E\right) < E\left(H_E^*\right)$ as $\psi'(\theta_i) < 1$ for $i = 1, \ldots, k$. Now, let $\hat{\theta}_i = \psi^{-1}(d_i)$ for $i = 1, \ldots, k$ (ψ is invertible when $\theta_i > 1 + \sqrt{\gamma}$). Further, let $\hat{\Theta}_k, \hat{\Psi}_k$, and $\hat{\Psi}_k'$ denote the diagonal matrices similarly constructed as Θ_k, Ψ_k, and Ψ_k', with θ_is replaced by $\hat{\theta}_i$s. Then, $\hat{\Theta}_k \xrightarrow{p} \Theta_k$,$\hat{\Psi}_k \xrightarrow{p} \Psi_k$, and $\hat{\Psi}_k' \xrightarrow{p} \Psi_k'$ element-wise.

Now, let $\hat{\eta}^* = \left(X^{*T}X^*\right)^{-1}X^{*T}Y$ be the least squares estimator from model (3.6). Then, $E(\hat{\eta}^*) = \left(X^{*T}X^*\right)^{-1}X^{*T}\tilde{X}\tilde{\eta} \xrightarrow{p} \left(\Theta_k\Psi_k'\Psi_k^{-1}\right)^{1/2}\tilde{\eta}$ under the true model (3.5).

Then, $\hat{\eta} = \left(\hat{\Theta}_k \hat{\Psi}'_k \hat{\Psi}_k^{-1}\right)^{-1/2} \hat{\eta}^*$ is asymptotically unbiased for $\tilde{\eta}$, and

$$\begin{aligned} E\left(\hat{\eta}^T \hat{\Theta}_k \hat{\eta}\right) &= E\left[Y^T X^* \left(X^{*T} X^*\right)^{-1} \left(\hat{\Psi}_k^{-1} \hat{\Psi}'_k\right)^{-1} \left(X^{*T} X^*\right)^{-1} X^{*T} Y\right] \\ &= tr\left[\left(X^{*T} X^*\right)^{-1} \left(\hat{\Psi}_k^{-1} \hat{\Psi}'_k\right)^{-1}\right] \\ &\qquad + \tilde{\eta}^T \left(\tilde{X}^T X^*\right)^2 \left(X^{*T} X^*\right)^{-2} \left(\hat{\Psi}_k^{-1} \hat{\Psi}'_k\right)^{-1} \tilde{\eta} \\ &\xrightarrow{p} \tilde{\eta}^T \Theta_k \tilde{\eta} \quad \text{as } n \to \infty. \end{aligned}$$

Therefore, $\hat{\eta}^T \hat{\Theta}_k \hat{\eta}$ is an asymptotically unbiased estimator for $E\left(\tilde{H}_E\right)$. Using this estimator, we can further estimate the proportion of variability (R^2) in Y explained by X in model (3.5) given by $\hat{R}^2 = \hat{\eta}^T \hat{\Theta}_k \hat{\eta} / H_T$, where $H_T = n^{-1} Y^T \left(I_n - J_n/n\right) Y$ denotes the total variance in the observed outcomes.

3.4.1 Implications When k^* is Incorrectly Estimated, or When $k^* < k$

Our previous derivation assumes $k = k^*$ and k^* is correctly estimated. However, this assumption may not always be satisfied in real data. Let the estimate for k^* be $\hat{k}^*$, and $c = \min\left(k^*, \hat{k}^*\right)$. As long as $\hat{k}^*$ is finite, $\hat{\Theta}_c, \hat{\Psi}_c$, and $\hat{\Psi}'_c$ will remain consistent to Θ_c, Ψ_c, and Ψ'_c, respectively. Moreover, from Theorem 3.2, as $v_r^T u_i \xrightarrow{p} 0$ for $r = k^* + 1, \ldots, p$,

$$X^{*T} \tilde{X}/n \xrightarrow{p} \begin{bmatrix} \Theta_c \Psi'_c \Psi_c & 0 \\ 0 & 0 \end{bmatrix}_{\hat{k}^* \times k}.$$

In the previous matrix notation, zeroes are augmented as required to achieve the specified matrix dimensions. Then,

$$\begin{aligned}
\hat{\eta} &= \left(\hat{\Theta}_{\hat{k}^*}\hat{\Psi}'_{\hat{k}^*}\hat{\Psi}^{-1}_{\hat{k}^*}\right)^{-1/2}\left(X^{*T}X^*\right)^{-1}X^{*T}Y \\
&\xrightarrow{p} \begin{bmatrix} I_c & 0 \\ 0 & 0 \end{bmatrix}_{\hat{k}^*\times k} \tilde{\eta}, \quad \text{and} \\
E\left(\hat{\eta}^T\hat{\Theta}_{\hat{k}^*}\hat{\eta}\right) &= E\left[Y^TX^*\left(X^{*T}X^*\right)^{-1}\left(\hat{\Psi}^{-1}_{\hat{k}^*}\hat{\Psi}'_{\hat{k}^*}\right)^{-1}\left(X^{*T}X^*\right)^{-1}X^{*T}Y\right] \\
&= tr\left[\left(X^{*T}X^*\right)^{-1}\left(\hat{\Psi}^{-1}_{\hat{k}^*}\hat{\Psi}'_{\hat{k}^*}\right)^{-1}\right] \\
&\qquad + \tilde{\eta}^T\left(\tilde{X}^TX^*\right)\left(X^{*T}X^*\right)^{-2}\left(\hat{\Psi}^{-1}_{\hat{k}^*}\hat{\Psi}'_{\hat{k}^*}\right)^{-1}\left(X^{*T}\tilde{X}\right)\tilde{\eta} \\
&\xrightarrow{p} \tilde{\eta}^T\begin{bmatrix} \Theta_c & 0 \\ 0 & 0 \end{bmatrix}_{k\times k}\tilde{\eta} \quad \text{as } n\to\infty
\end{aligned}$$

Therefore, the asymptotic bias in the estimate $\hat{\eta}^T\hat{\Theta}_{\hat{k}^*}\hat{\eta}$ for $E\left(\tilde{H}_E\right)$ will be $-\sum_{i=c+1}^{k}\theta_i\tilde{\eta}_i^2$, where $\tilde{\eta}_i$ is the i^{th} element of $\tilde{\eta}$. When $k^*=k$, and k^* is correctly or over-estimated, this result implies that $\hat{\eta}^T\hat{\Theta}_{\hat{k}^*}\hat{\eta}$ is asymptotically unbiased for $E\left(\tilde{H}_E\right)$. However, if $k^*<k$, or k^* is under-estimated, then $\hat{\eta}^T\hat{\Theta}_{\hat{k}^*}\hat{\eta}$ will asymptotically under-estimate $E\left(\tilde{H}_E\right)$.

3.5 Adjusting the Shrinkage Bias to Improve Prediction Accuracy

Lee et al. (2010) showed that in the high-dimensional setting, if the sample eigenvectors are used to predict the PC scores of new observations, then the predicted PC scores are biased towards zero. This shrinkage phenomenon can result in loss of prediction accuracy in high-dimensional PLS as well, since we are using the sample PC scores to estimate the regression parameters. As the predicted PC scores of the

new observations are shrunk towards zero, they need to be shrinkage-adjusted first, before using them for prediction. The shrinkage adjustment procedure is a direct applicaton of the following result (*Lee et al.*, 2010),

Result 3.1. *If* $\theta_i > 1 + \sqrt{\gamma}$,

$$\sqrt{\frac{E\left(z_{new,i}^2\right)}{E\left(z_{ji}^2\right)}} \xrightarrow{p} \rho(\theta_i) = \frac{\theta_i - 1}{\theta_i - 1 + \gamma},$$

where $z_{new,i} = x_{new}^T v_i$, $z_{ji} = x_j^T v_i$, x_{new} *is a new observation coming from the same distribution as the observations in* X, *and* x_j *is the* j^{th} *row of* X, *for any* $j = 1, \ldots, n$.

Then, we can adjust the predicted scores using the predicted shrinkage factor $\rho(\hat{\theta}_i)$s, where $\hat{\theta}_i = \psi^{-1}(d_i)$ is the consistent estimator for θ_i. Then the shrinkage-adjusted predicted score is given by $z_{adj} = z_{new,i}/\rho(\hat{\theta}_i)$. Note that, the above result is assuming that $\sigma_x^2 = 1$. If $\sigma_x^2 \neq 1$, then the shrinkage factors can be estimated using the algorithm proposed by *Lee et al.* (2010) (Section 2.4), or using Theorem 2.4 in Chapter II (under spiked population model, both methods provide almost identical shrinkage factor estimates, as discussed in Chapter II).

3.6 Numerical Simulations

We performed extensive simulation studies to compare the performance of our proposed TPLS method with the traditional PLS and SPLS methods, both in terms of the R^2 and mean-squared error of prediction (MSEP). We simulated from the following model,

$$\begin{aligned} X_{n\times p} &= TP^T + E \\ Y_{n\times 1} &= Tg + F, \end{aligned} \tag{3.8}$$

with one outcome, $n = 500$ samples, and $p = 10000$ predictors, $\gamma = p/n = 20$. We first generated the rows of $T_{n\times k}$ from $N(0, \Lambda)$ independently, where $k = 10$ and Λ is diagonal with diagonal elements $\lambda_i = 65 - 5i$ for $i = 1, \ldots, k$. Then, we considered three sparsity levels in X: 99% sparse ($n_s = 10$), 90% sparse ($n_s = 100$), and non-sparse ($n_s = 1000$), and selected $P^T = (n_s)^{-1/2} \left[\underbrace{I_k \quad \ldots \quad I_k}_{n_s} \quad \underbrace{0 \quad \ldots \quad 0}_{p-n_s} \right]$ accordingly. The elements of E were simulated i.i.d. from $N(0, \sigma_x^2)$. Then, the population eigenvalues of X are given by $\tau_i = \lambda_i + \sigma_x^2$ for $i \leq k$, and $\tau_i = \sigma_x^2$ for $i > k$. To explore different signal-to-noise ratios, we considered three choices for $\sigma_x^2 = 1, 2, 5$. $\sigma_x^2 = 1$ and $\sigma_x^2 = 2$ implies that compared to the non-spikes, the population spikes of X are very large or moderately large, respectively. $\sigma_x^2 = 5$ implies the spikes are close to the non-spikes. In fact, when $\sigma_x^2 = 5$, the last three spikes $\tau_8 = 25, \tau_9 = 20$ and $\tau_{10} = 15$ are smaller than $\sigma_x^2 + \sigma_x^2\sqrt{\gamma} = 27.36$. Therefore, in this case, due to the presence of close spikes, the estimate for R^2 given by the method described in Section 3.4, should underestimate the true R^2. Next, to simulate Y, we selected the all elements of $g = (g_1, \ldots, g_k)$ to be equal to unity, and generated $F \sim N(0, \sigma_y^2 I_n)$. We considered three values of $R^2 = 0.1, 0.5$, and 0.7, and selected $\sigma_y^2 = [(1 - R^2)/R^2] \left(\sum_{i=1}^k \lambda_i g_i \right)$ accordingly.

We applied the traditional PLS, SPLS and our proposed TPLS method to fit the simulated data for each choice of sparsity level, σ_x^2, and R^2. For all methods, the number of components (and the thresholding parameter for SPLS) were selected based on 10-fold cross-validation (CV). For the TPLS method, the number of distant spikes k^* was estimated using the algorithm proposed by *Lee et al.* (2010) (Section 2.4). After fitting the model with different methods, we calculated the observed coefficient of determination $\hat{R}^2$ for each of them. Using the parameter estimates from the TPLS method, we further calculated the adjusted R^2 estimator $\hat{R}^2_{adj}$ by applying the method described in Section 3.4.

To assess the prediction performances of different methods, we simulated a test

dataset with $n = 500$ samples, using the same model as described above, and predicted the outcomes based on the parameter estimates obtained from different methods. For TPLS, we recorded both the predicted outcomes with and without the shrinkage adjustment of the predicted PC scores. Then, we calculated the MSEP for each method, scaled by the variance of the outcomes in the test dataset.

We simulated 100 training and test datasets for each of the different sparsity levels, σ_x^2, and R^2, using the simulation method described above. The box plots for the observed R^2s are presented in Figures 3.2, 3.3, and 3.4. We can see in all of the scenarios, the traditional PLS provided R^2 very close to one even when the true R^2 is as small as 0.1, which is a clear indication of over-fitting. Except for the cases with 99% sparsity level, the SPLS method also provided very high R^2 regardless of the true R^2. Even for cases with 99% sparsity level, the observed R^2s from SPLS were not stable, in the sense that, they were often as high as one or as low as the true R^2. On the other hand, the TPLS method provided stable observed R^2s which were almost identical to the true R^2 when $\sigma_x^2 = 1$ or when $R^2 = 0.1$. When $\sigma_x^2 = 2$, the observed R^2s from the TPLS method were almost identical to the true R^2s when $R^2 = 0.1$ or 0.5, and slightly smaller than the true R^2 when $R^2 = 0.7$. When $\sigma_x^2 = 5$, due to the presence of three close spikes, the TPLS method under-fitted the model (as all sample eigenvectors corresponding to nonzero sample eigenvalues were asymptotically orthogonal to the population eigenvectors corresponding to the close spikes), and the observed R^2s were moderately smaller than the true R^2 for $R^2 = 0.5$ or 0.7. The averages of the adjusted R^2s (denoted by the blue horizontal lines) also followed similar patterns. The behavior of the adjusted R^2s can be explained by k^*, and its estimate $\hat{k}^*$. When $\sigma_x^2 = 1$ or 2, $k^* = k = 10$, and when $\sigma_x^2 = 5$, $k^* = 7 < k$. From Table 3.1, we can see that $\hat{k}^* = 10$ for all of the simulated datasets when $\sigma_x^2 = 1$. Therefore, the adjusted R^2 remained asymptotically unbiased to the true R^2 in this case. When $\sigma_x^2 = 2$, $\hat{k}^* = 10$ for 82% of the simulations, and 9 for the rest, which

resulted in slight under-estimation of R^2. For $\sigma_x^2 = 5$, because the condition that $k^* = k$ was clearly violated, the adjusted R^2 asymptotically under-estimated the true R^2. Therefore, the simulation results overall do not show any indication of over-fitting in the TPLS method, rather it can under-fit the model in some scenarios, especially when σ_x^2 is large and thus some of the population spikes in the predictors are very close to the non-spikes.

However, this under-fitting in TPLS is still preferable than the over-fitting in traditional PLS and SPLS methods in terms of prediction accuracy, as shown in the box plots for the MSEPs (Figures 3.5, 3.6, and 3.7). The TPLS method with the shrinkage adjustment provided lower median MSEPs than the other methods across almost all the different scenarios. Only for the case with 99% sparsity, $R^2 = 0.5$ and $\sigma_x^2 = 5$, or the cases with 99% sparsity and $R^2 = 0.7$, SPLS performed the best. Across all the scenarios, the TPLS method with shrinkage adjustment provided lower median MSEP than without the shrinkage adjustment, which emphasizes the importance of the shrinkage adjustment.

Overall, these results suggest that the worst case scenario for our proposed TPLS method, in terms of under-fitting the model, is when close spikes are present in the covariance matrix of the predictors. However, even in this worst case scenario, it shows better prediction performance than the traditional PLS method across all sparsity levels, and also better than the SPLS method when the sparsity level is not extremely high.

σ_x^2	**Estimated no. of distant spikes**					
	10	9	8	7	6	5
1	100					
2	82	18				
5			1.33	50	47.67	1

Table 3.1: Percentage of training datasets where the number of distant spikes were estimated to be between five and ten.

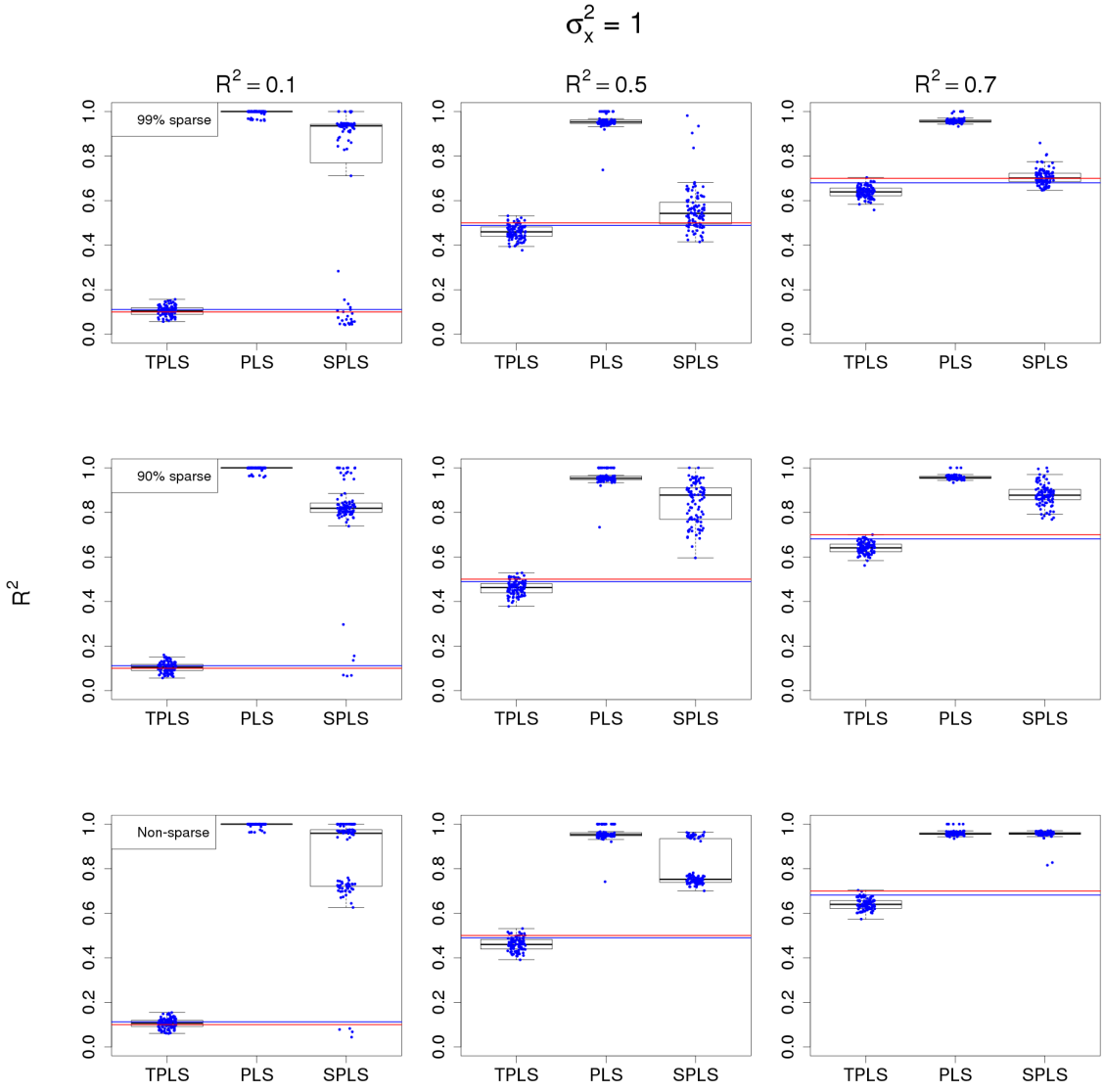


Figure 3.2: Observed R^2s for TPLS, PLS, and SPLS methods when the spikes are much larger the non-spikes, i.e, $\sigma_x^2 = 1$. The red horizontal line shows the true R^2, and the blue horizontal line shows the averages of the adjusted R^2 estimates.

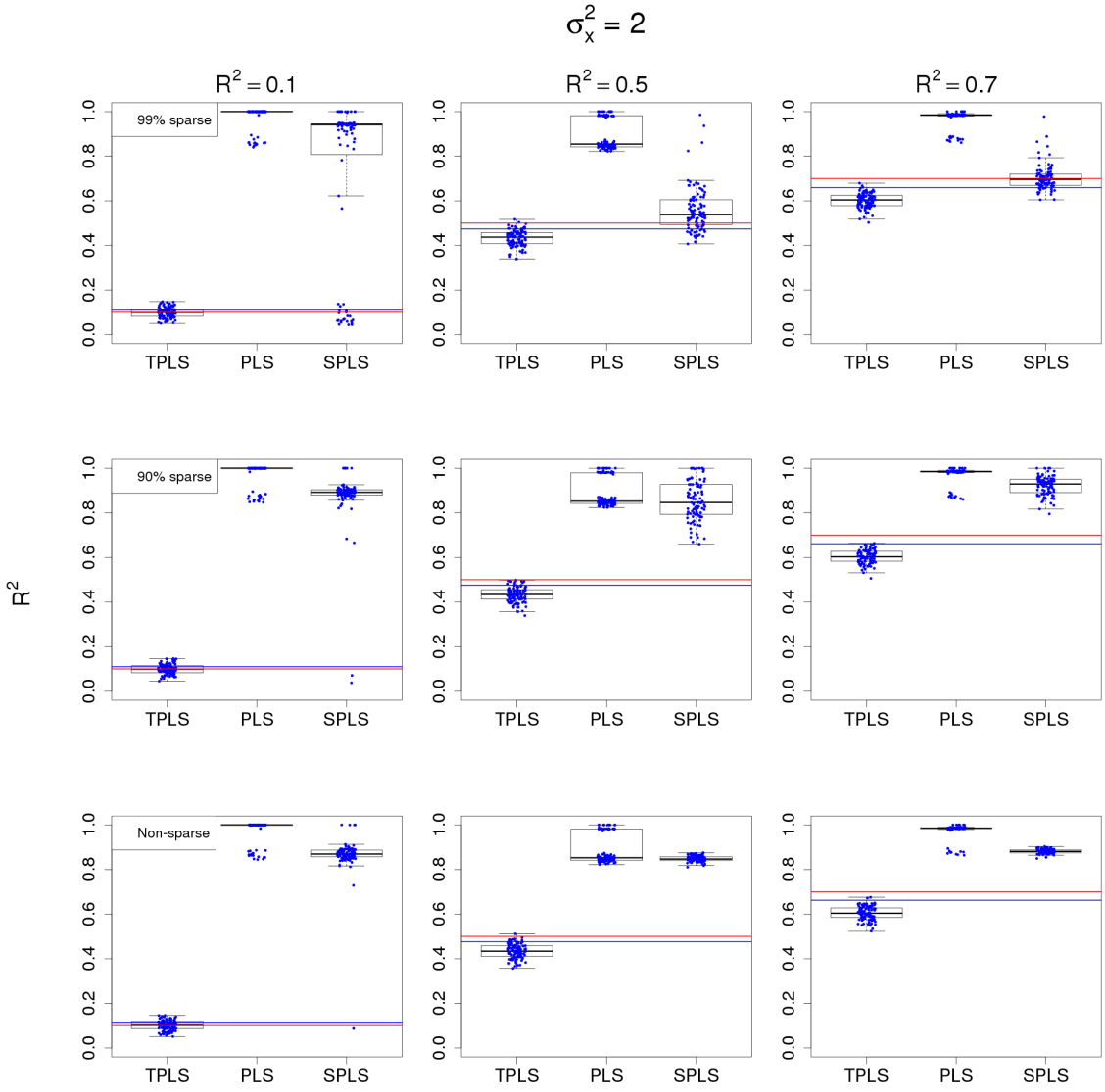


Figure 3.3: Observed R^2s for TPLS, PLS, and SPLS methods when the spikes are moderately large compared to the non-spikes, i.e, $\sigma_x^2 = 2$. The red horizontal line shows the true R^2, and the blue horizontal line shows the averages of the adjusted R^2 estimates.

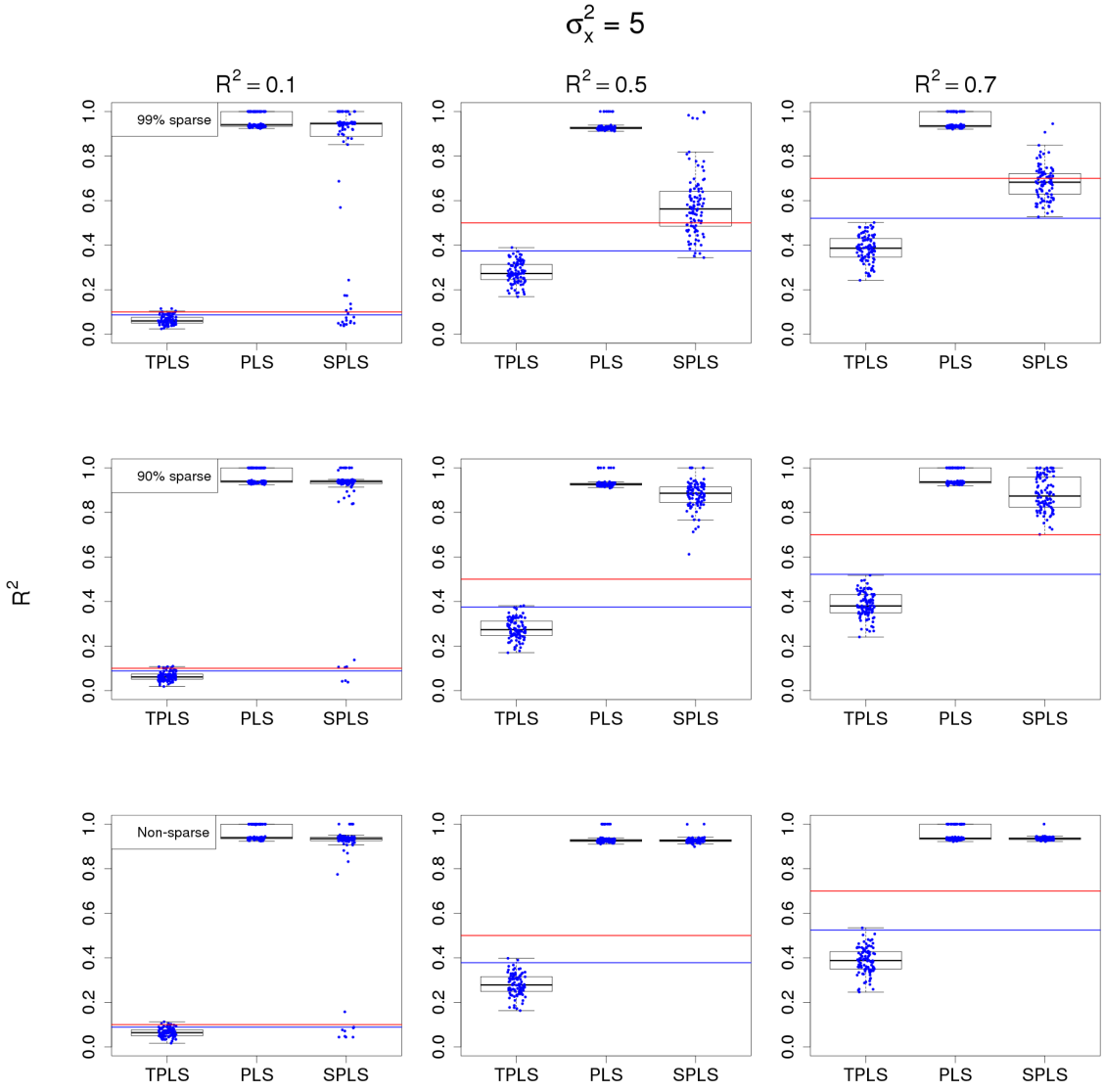


Figure 3.4: Observed R^2s for TPLS, PLS, and SPLS methods when the spikes are close to the non-spikes, i.e, $\sigma_x^2 = 5$. The red horizontal line shows the true R^2, and the blue horizontal line shows the averages of the adjusted R^2 estimates.

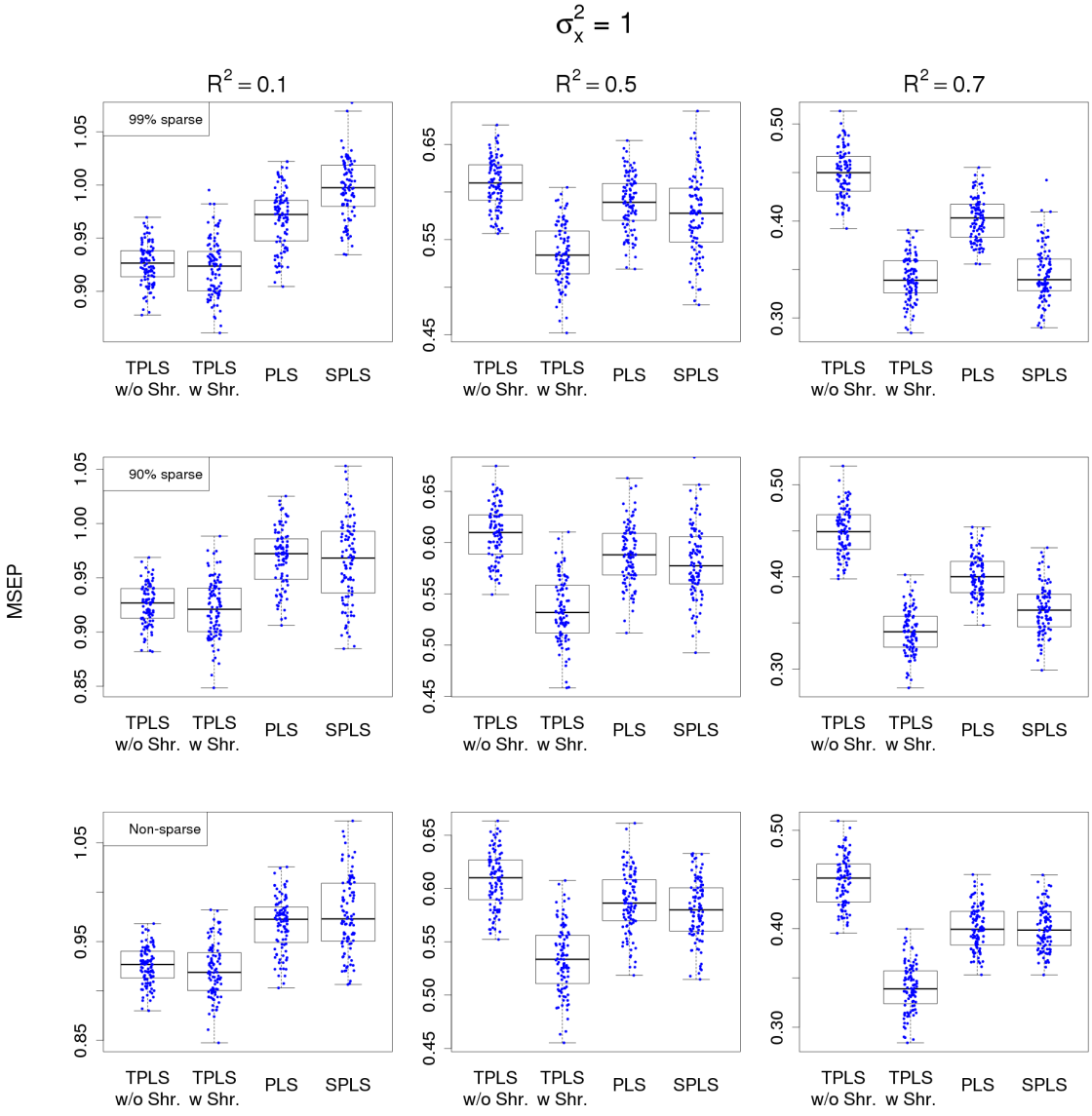


Figure 3.5: MSEP for TPLS, PLS, and SPLS methods when the spikes are much larger the non-spikes, i.e, $\sigma_x^2 = 1$. The MSEPs are scaled by the variance of the observed outcomes.

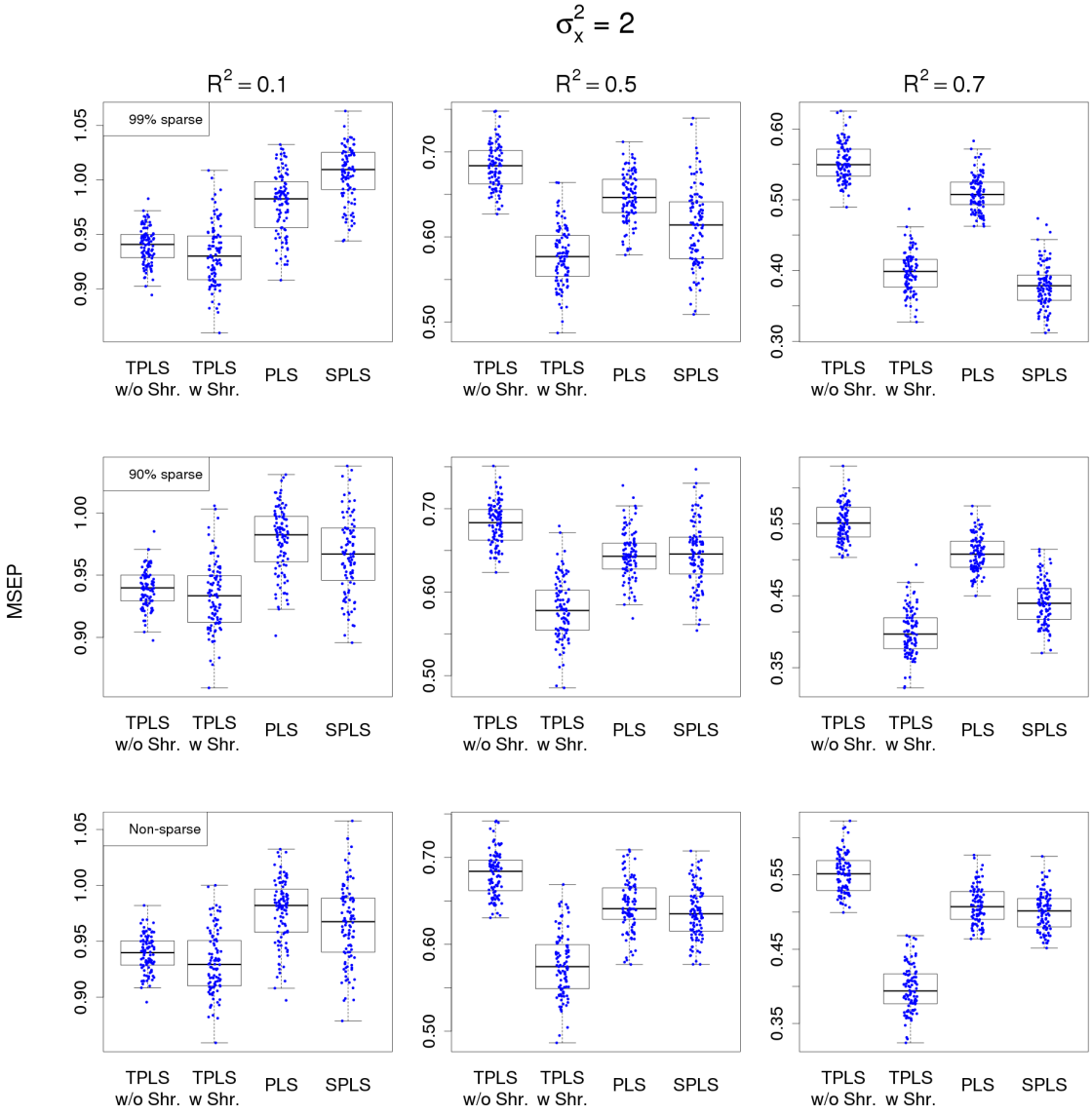


Figure 3.6: MSEP for TPLS, PLS, and SPLS methods when the spikes are moderately large compared to the non-spikes, i.e, $\sigma_x^2 = 2$. The MSEPs are scaled by the variance of the observed outcomes.

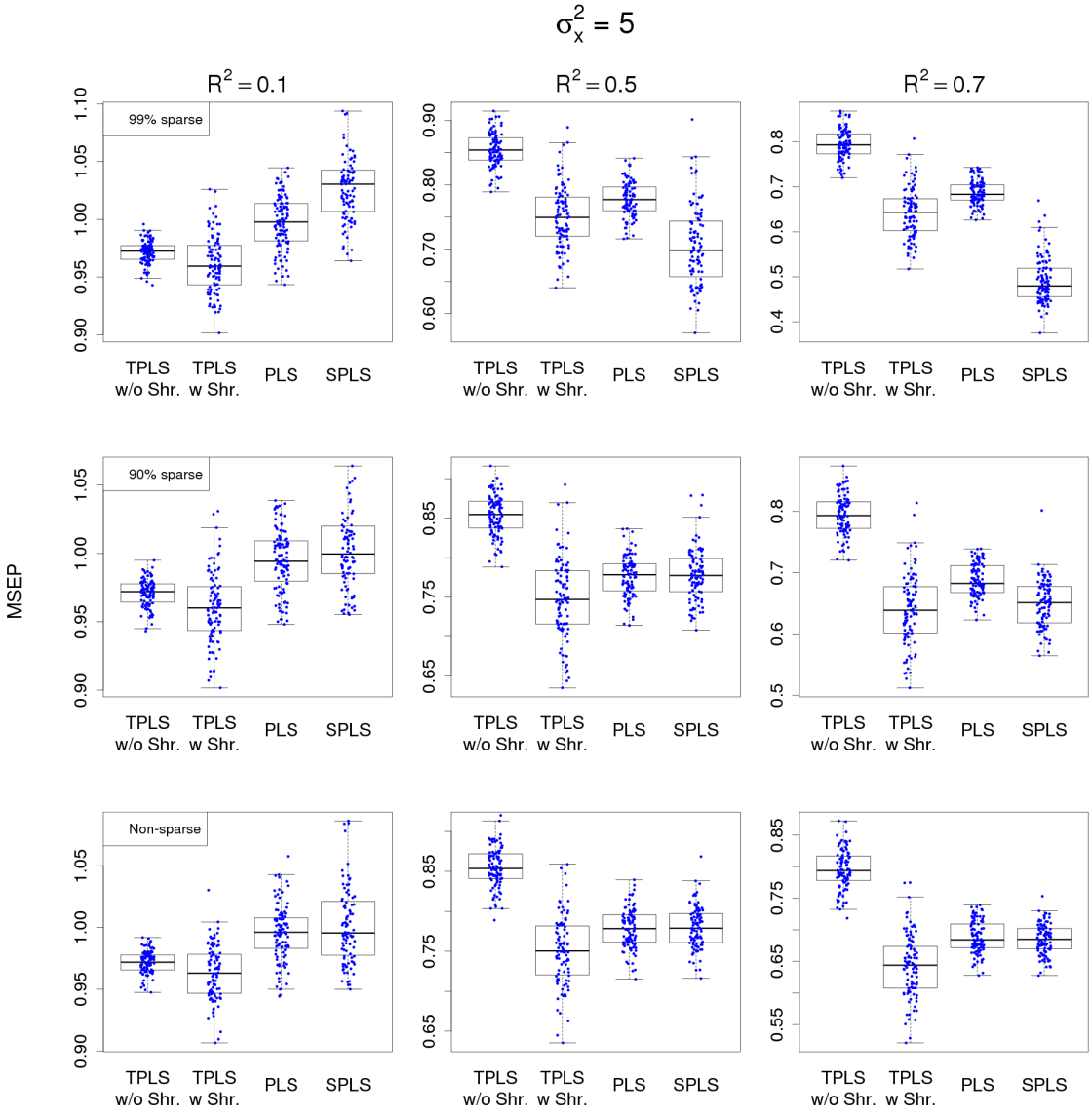


Figure 3.7: MSEP for TPLS, PLS, and SPLS methods when the spikes are close to the non-spikes, i.e, $\sigma_x^2 = 5$. The MSEPs are scaled by the variance of the observed outcomes.

3.7 ADNI Data Example

We applied the proposed TPLS method along with the traditional PLS and SPLS methods on the Alzheimer's Disease Neuroimaging Initiative (ADNI) data to compare their performances in real world problems. Our goal was to model the monthly decline rates in composite scores for memory (MEM) and composite scores for executive functioning (EF) jointly, based on the cortical thickness measurements at baseline across the brain (*Crane et al.*, 2012; *Gibbons et al.*, 2012). For this purpose, we used 825 samples with early (EMCI) and late mild cognitive impairment (LMCI) diagnosed at baseline. We calculated their monthly decline rates in MEM and in EF scores by subtracting the scores recorded on their last visit, from the scores recorded at baseline, and dividing them by the durations between their first and last visit. The rate of decline in MEM and EF scores can help us understand which regions of the brain surface are associated with the cognitive decline rate, which ultimately can result in the Alzheimer's disease (AD).

Pre-processed magnetic resonance imaging (MRI) scans at baseline were downloaded from the ADNI data repository (`http://www.loni.usc.edu/ADNI/`). T1-weighted brain MRI scans at baseline were acquired using a sagittal 3D MP-RAGE sequence following the ADNI MRI protocol (Jack et al., 2010; Jack et al., 2008). MRI scans were processed prior to download as previously described (*Jack et al.*, 2010, 2008). As detailed in previous studies, FreeSurfer V5.1, a widely employed automated MRI analysis approach, was used to process MRI scans and extract cortical thickness determined by automated segmentation and parcellation (*Kim et al.*, 2013; *Nho et al.*, 2013, 2015). The cortical surface was reconstructed to measure thickness at each vertex using cognitively normal adult participants. The cortical thickness was calculated by taking the Euclidean distance between the grey/white boundary and the grey/cerebrospinal fluid (CSF) boundary at each vertex on the surface (*Chung et al.*, 2010; *Dale et al.*, 1999; *Fischl et al.*, 1999). Prior to fitting the PLS models,

we first regressed out the possible confounders (sex, scan type, intracranial volume, age, education, baseline MEM and EF scores, and an indicator whether they were diagnosed at the EMCI or LMCI stage) from both the outcomes and the cortical thickness measurements. We also centered and scaled the cortical thickness measurements so that the standard deviation for measurements on each vertex across all the samples becomes unity. For the traditional PLS method, we first used 10-fold CV to select the number of PLS components (selected to be two). Then we fitted the model with two PLS components and the red point on the right hand panel of Figure 3.8 shows the observed R^2 ($\hat{R}^2 = 0.1605$) for that fit. We further investigated the effect of over-specifying the number of PLS components by varying the number of selected components from two to 25 and calculating the observed R^2 for each of them. The observed R^2s are presented as blue dots in the right-hand panel of Figure 3.8. The results show that the traditional PLS method does not provide stable observed R^2s and the observed R^2s can vary between zero and unity, with the median being close to unity, which potentially indicates over-fitting. For the TPLS method also, we first estimated the number of distant spikes ($\hat{k}^* = 19$) in the cortical thickness measurements, and selected the number of PLS components (selected to be four) using 10-fold CV. Then we fitted the model using TPLS with 19 PCs and four PLS components, and the red point on the left hand panel of Figure 3.8 shows the observed R^2 ($\hat{R}^2 = 0.0696$) for that fit. We also investigated the stability of the TPLS fit, specifically whether it over-fits if more PC or PLS components are included in the model, by fitting the model with the number of PCs varying from 19 to 25, and the number of PLS components varying from four to 25. The resulting observed R^2s are presented on the left-hand panel of Figure 3.8. We also calculated the adjusted R^2 estimate using 19 sample PCs, and it is represented by the blue horizontal line in the plot. The results show that, the observed R^2s for TPLS are stable and very close to the adjusted R^2 estimate, $\hat{R}^2_{adj} = 0.0758$. The SPLS method selected the thresholding

parameter to be zero through 10-fold CV, which implies that the results from SPLS are identical to that from the traditional PLS method, and hence those were omitted from the plot.

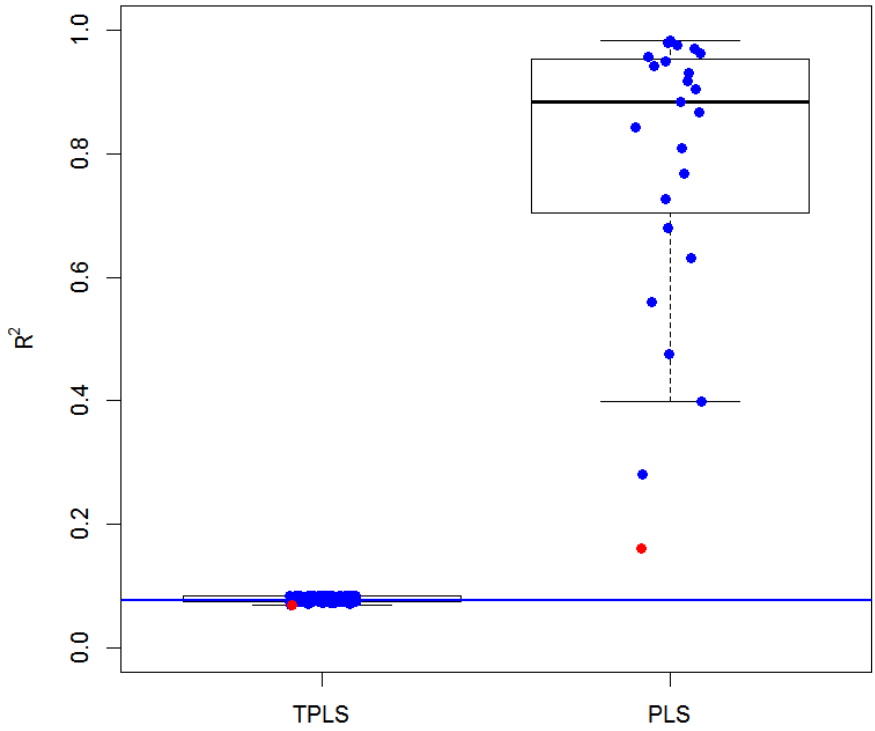


Figure 3.8: Observed R^2s for TPLS and PLS methods for different specifications of the number of components and PCs (for TPLS only). Number of PLS components for traditional PLS varies from 2 to 25, and number of PCs and PLS components for TPLS varies from 19 to 25, and from 4 to 25, respectively. The red point on the left hand panel shows the observed R^2 at 19 PCs and 4 PLS components (selected by 10-fold CV). The red point on the right hand panel shows the observed R^2 at 2 PLS components (selected by 10-fold CV). The blue horizontal line shows the adjusted R^2 estimate.

We further mapped the regression coefficients ($\hat{B}$) on the brain surface for both PLS and TPLS methods using different number of PLS components. We mapped $\hat{B}$s for the PLS method with two PLS components, and for the TPLS method with

19 PCs and four PLS components (selected by 10-fold CV as mentioned earlier). In addition, we mapped $\hat{B}$s for both methods with 10 and 15 PLS components, in order to investigate the effect of over-specifying the number of PLS components. Figures 3.9 and 3.10 show that for both the outcomes, PLS with two components provided almost homogeneous plots with no specific regions clearly highlighted as having an effect on the outcomes. When used 10 or 15 PLS components, the brain surface maps became substantially different from the maps using two components. The vertices with strong effects on the outcomes, were spread across the entire brain surface, and they did not form any contiguous meaningful region. On the other hand, the regression coefficient estimates from TPLS (Figures 3.11 and 3.12) remained robust against the selection of different number of PLS components. The vertices that were shown to have strong effects on the outcomes, formed contiguous and meaningful regions on the brain surface. In particular, decreased cortical thickness in the bilateral frontal, parietal and medial temporal lobes including the entorhinal cortex was shown to have a hastening effect on the decline rate in the MEM score. The medial temporal lobe (MTL) including the entorhinal cortex is the first region to show AD-related neurodegeneration, and the decline rate in memory performance has been shown to be associated with MTL atrophy rates in people at-risk for AD (*Braak and Braak*, 1996; *Nho et al.*, 2012; *Fox et al.*, 1996; *Jagust et al.*, 2006; *Rusinek et al.*, 2003). We can also see the decline rate in the EF score to be impacted by decreased cortical thickness in the bilateral temporal, parietal and especially the frontal lobes, which are known to be important for executive functions. Previous studies have shown strong associations between executive function deficits and brain atrophy in regions of the frontal, parietal, and temporal lobes (*Nho et al.*, 2012; *Pa et al.*, 2010; *Huey et al.*, 2009; *Thomann et al.*, 2008).

To investigate the effects of different methods on the MSEP, we divided the samples randomly in training and test samples: 413 randomly selected samples were

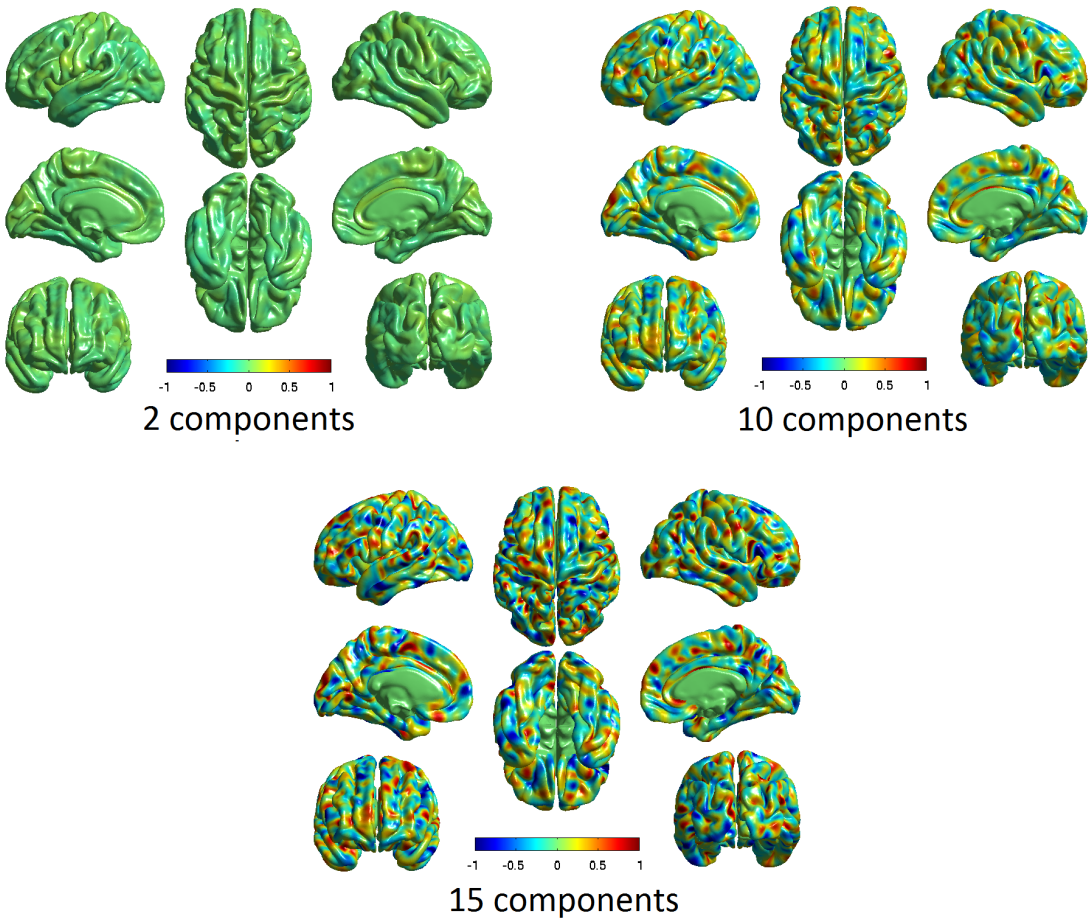


Figure 3.9: PLS regression coefficient estimates corresponding to the MEM scores mapped on the brain surface.

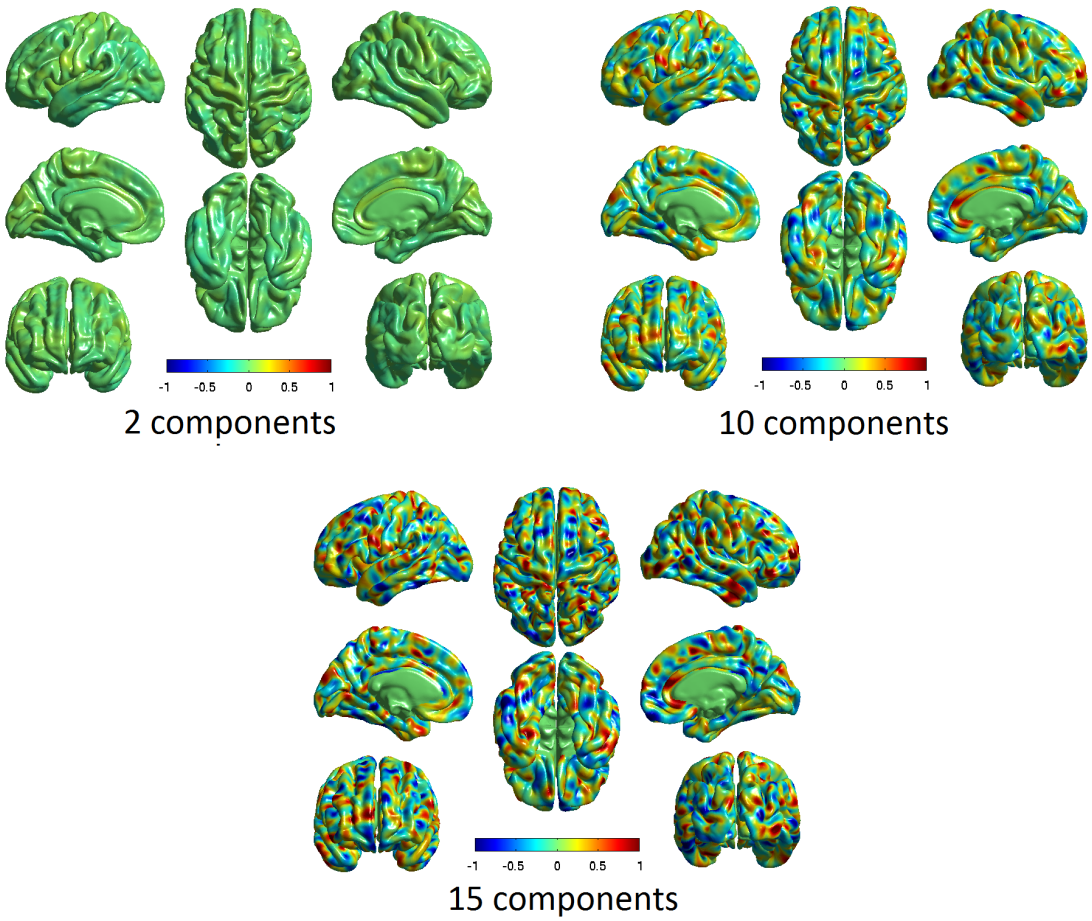


Figure 3.10: PLS regression coefficient estimates corresponding to the EF scores mapped on the brain surface.

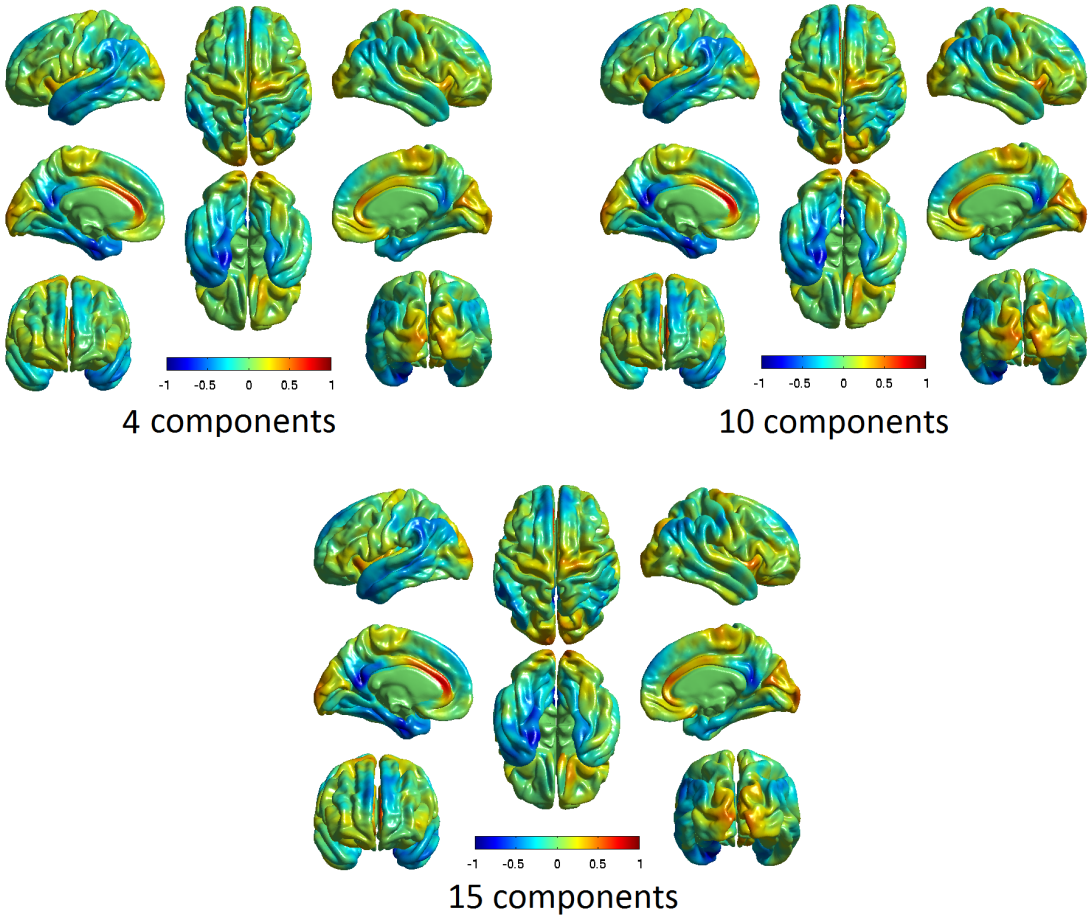


Figure 3.11: TPLS regression coefficient estimates corresponding to the MEM scores mapped on the brain surface.

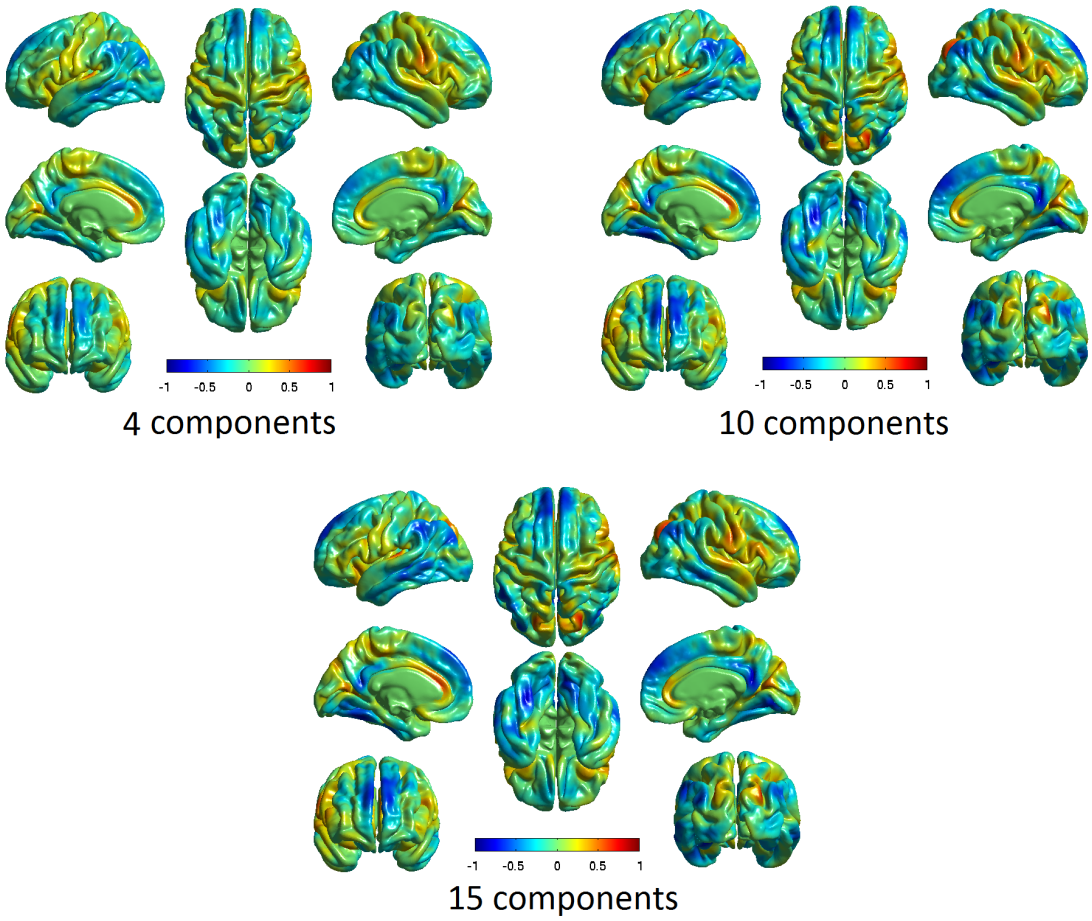


Figure 3.12: TPLS regression coefficient estimates corresponding to the EF scores mapped on the brain surface.

considered as training samples, and rest of the 412 samples as test samples. On the training samples, we applied the traditional PLS method using the number of PLS components selected via 10-fold CV. To apply the TPLS method, we first calculated the number of spiked PCs for the cortical thickness measurements, and selected the the number of PLS components via 10-fold CV. Using the estimated model parameters, we predicted the outcomes in the test data, and calculated the MSEPs. We performed this analysis 50 times using different selection of training and test samples. In all of the iterations, the SPLS method, similar to the case where all 825 samples were included in the model, provided identical results as the traditional PLS, and hence we omitted it from our results. The observed R^2s from the training data, and the MSEPs (scaled by the sum of variances of the columns of Y) from the test data are presented in Figures 3.13 and 3.14 respectively. The results suggest that the observed R^2s for the traditional PLS are unstable and are spread out between zero and unity, even when the number of components are selected by CV. The median observed R^2 is 0.3086, which is much larger than the observed R^2 resulted from fitting the model on all 825 samples with number of PLS components selected by 10-fold CV (as discussed earlier in this section). This emphasizes the unreliability of the observed R^2 values obtained from the traditional PLS method, and the risk of falsely inferring that the predictors account for a higher proportion of variability in the outcomes, when the high R^2 observation is possibly due to over-fitting. On the other hand, our proposed TPLS method provides stable observed R^2s, and the median observed $R^2 = 0.0892$ is very close to the observed R^2 when all of the 825 samples were included, and the TPLS model is fitted with 19 PCs and four PLS components (selected by 10-fold CV). We further calculated the adjusted R^2 estimates for each selection of training samples, and the median adjusted R^2 estimate ($\hat{R}^2_{adj} = 0.095$) is represented by the blue horizontal line in Figure 3.13, which is also very close to the median observed R^2 from TPLS.

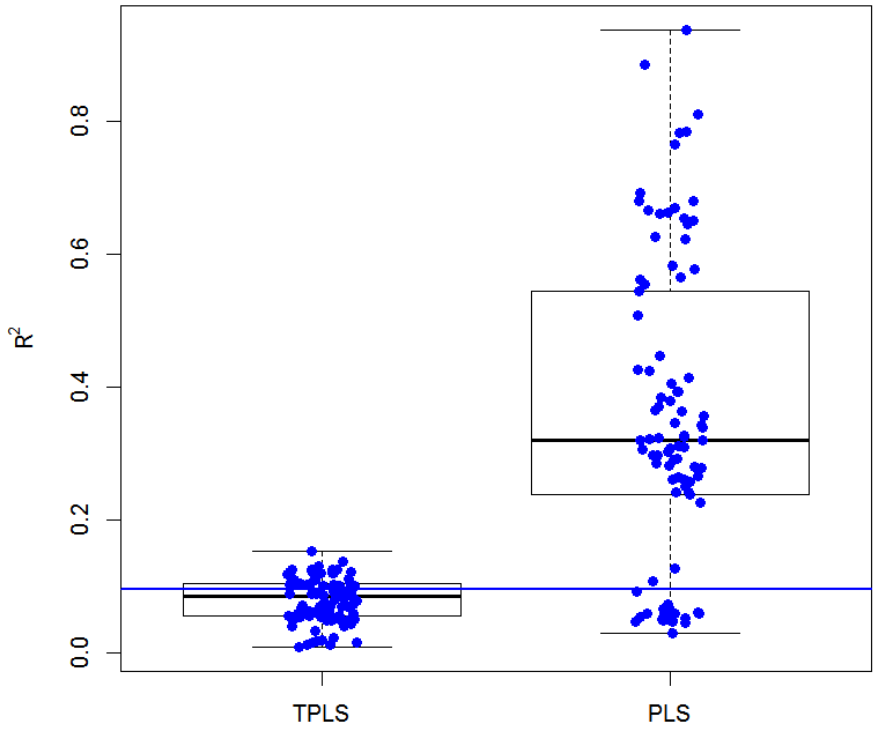


Figure 3.13: Observed R^2s for TPLS and PLS methods for 50 randomly selected training sample sets. Number of PLS components for both methods were selected using a 10-fold CV. The blue horizontal line shows the median adjusted R^2 estimate using the method described in Section 3.4.

Figure 3.14 shows that the median MSEPs for our proposed TPLS method (0.8136 without shrinkage adjustment, and 0.8293 with shrinkage adjustment) are lower than the median MSEP for the traditional PLS method (0.9086), and the spread of MSEPs is also lower for the TPLS method. Even though the median MSEP is slightly larger for the shrinkage adjusted TPLS method compared to the shrinkage unadjusted version, the observed MSEPs and the spread of the MSEPs are almost identical between these two versions of TPLS, which indicates that the effect of the shrinkage phenomenon is negligible in this data.

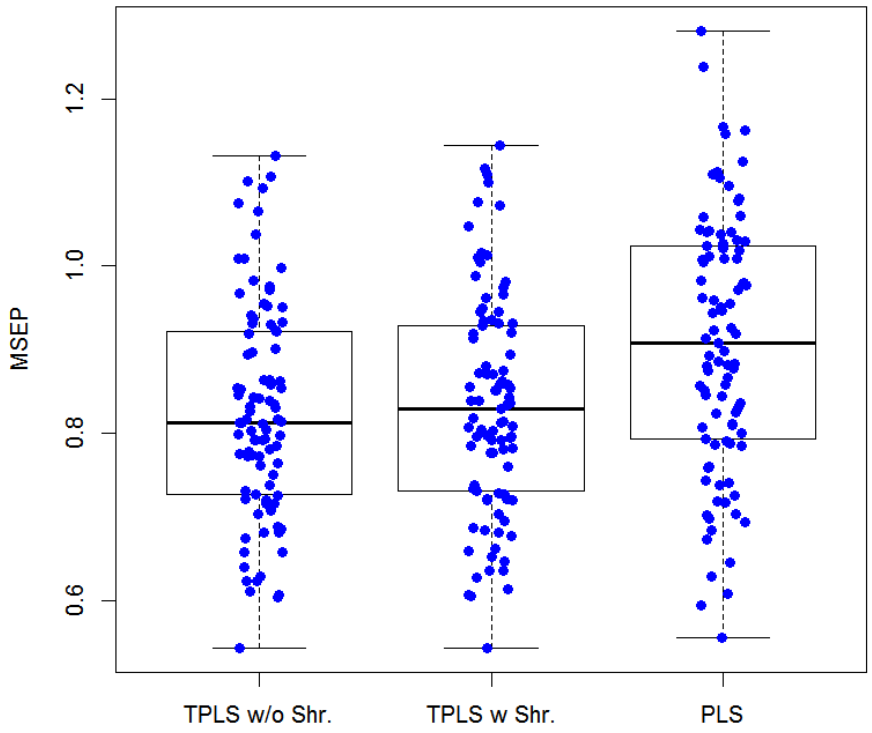


Figure 3.14: MSEP for TPLS and PLS methods for 50 randomly selected training and test sample sets. The MSEPs are scaled by the sum of the column-wise variances of the outcomes.

In summary, this ADNI data example shows that the observed R^2s from the traditional PLS, are unreliable, and can vary hugely depending on the number of PLS

components selected, even when the number of PLS components are selected using CV. On the other hand, the TPLS method provides stable model fits and observed R^2s across varying choice of number of PCs and number of PLS components. Moreover, TPLS performs better than the traditional PLS method in terms of prediction accuracy as well.

3.8 Discussion

In this chapter, we proposed a two-stage PLS (TPLS) method to address the over-fitting and shrinkage problems of PLS in high-dimensional data. Our method is robust and does not require any sparsity assumption or variable selection. We further provided a method to calculate the asymptotically unbiased estimator of the proportion of variablity in the outcomes that can be explained by the predictors. Through numerical simulations and real data applications, we evaluated the performance of the proposed method, as well as the traditional PLS and the SPLS methods. We showed that the traditional PLS and SPLS methods over-fit models with high-dimensional predictors under most of the scenarios, whereas the TPLS method protects against over-fitting by using a finite-dimensional subspace constraint spanned by the top sample eigenvectors. TPLS also performs best under most of the scenarios in terms of prediction accuracy among these methods. Only when the sparsity in the predictors is extremely high, and the predictors have a high effect on the outcomes, then SPLS provides more accurate predictions.

We noted that the worst case scenario for TPLS is when some of the population spikes are close to the non-spikes of the predictors, and those spikes have a high effect on the outcome. In such situations, TPLS may under-fit the model. However, even in its worst case scenarios, the prediction accuracy of TPLS is still better than the traditional PLS method, and except for the scenario with extremely sparse predictors, it is also better than the SPLS method.

We would also like to emphasize the unreliability in the traditional PLS estimates and the observed R^2 statistics. *Chun and Keleş* (2010) showed that the PLS estimates are not consistent in high-dimensional data. Due to the possibility of over-fitting, the R^2 statistic is also unreliable as seen from the numerical simulations and the ADNI data example. Therefore, even though PLS is an attractive method to analyze high-dimensional data, the researchers need to be careful when applying it, or may risk falsely attributing a higher effect of the predictors on the outcomes when the true effect might be much smaller.

Finally, we note that, even though our method is developed for linear models, similar modification of the PLS method can also be made to incorporate categorical outcomes. In the classification problems, the over-fitting problem has already been identified by other researchers (*Brereton and Lloyd*, 2014; *Gromski et al.*, 2015), and it can also affect the generalized linear model-based PLS methods (*Ding and Gentleman*, 2005; *Marx*, 1996; *Bastien et al.*, 2005). The extension of our method to the classification and generalized linear models, is left for future research.

CHAPTER IV

A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS

4.1 Introduction

Over the last decade, genome wide association studies (GWASs) have proved instrumental to unravelling the genetic complexities of hundreds of diseases and traits and their associations with common genomic variations. To date, thousands of GWASs have identified more than 4000 significant loci to be associated with human diseases and traits (*Welter et al.*, 2014). However, since most GWASs investigate a single disease or trait, they cannot exploit the cross-phenotype associations or pleiotropy (*Solovieff et al.*, 2013) where a single genetic variant can be associated with multiple phenotypes. Phenome-wide association study (PheWAS) has been proposed as an alternative approach to take advantage of the pleiotropy phenomenon by studying the impact of genetic variations across a broad spectrum of human phenotypes or 'phenome'. It is a complementary approach to GWAS in the sense that while GWAS attempts to identify phenotype-to-genotype associations, PheWAS uses a genotype-to-phenotype approach. The first PheWAS (*Denny et al.*, 2010) was published as a proof-of-principle study, which demonstrated that the PheWAS strategy could be applied to successfully identify the expected gene-disease associations. Additional

studies (*Denny et al.*, 2011; *Hebbring et al.*, 2013; *Ritchie et al.*, 2013; *Pendergrass et al.*, 2013; *Shameer et al.*, 2014) have shown that the PheWAS approach can further identify novel disease-SNP associations (*Hebbring*, 2014).

The PheWAS approach depends on the availability of detailed phenotypic information. Currently, most of the PheWASs are applied to clinical cohorts linked to electronic health records (EHR) and utilize the International Classification of Disease (ICD) billing codes to define clinical phenotypes. The ICD codes provide an intuitive ordering of the phenotypes based on clinical disease and trait classifications. Since the current genotyping and imputation technologies (*Marchini and Howie*, 2010) allow for genotyping tens of millions of variants at very low cost, an extensive PheWAS can attempt to investigate the genotype-phenotype associations by performing genome-wide association analyses in thousands of traits. We can interpret the PheWAS result of a single genetic variant by observing its associations across the phenome. Such a PheWAS is exhaustive in nature and has great potential to identify novel variants associated with clinical diseases. One of the main challenges of the PheWAS analysis is that most of the phenotypes are binary phenotypes with unbalanced ($1:5$) or often extremely unbalanced ($1:500$) case-control ratios (See Figure 4.1), since the data is collected in cohorts. Although standard asymptotic tests, such as the Wald, score and likelihood ratio tests, are relatively well calibrated and asymptotically equivalent (*Cox and Hinkley*, 1974) for common variants (minor allele frequency: MAF > 0.05) in balanced case-control studies, they can inflate type I error for low frequency ($0.01 <$ MAF ≤ 0.05) and rare variants (MAF ≤ 0.01) in unbalanced case-control studies (*Ma et al.*, 2013). Moreover, since the Wald and likelihood ratio tests need to calculate the likelihood or the maximum likelihood estimator under the full model, which is computationally expensive, they are not scalable for the amount of tests that PheWASs attempt. On the other hand, the score test is computationally efficient as it does not need to calculate the maximum likelihood

under the full model. However, as mentioned before, it suffers from having highly inflated type I error rates in unbalanced studies. *Ma et al.* proposed Firth's penalized likelihood ratio test (*Firth*, 1993) as a solution to control the type I error rates in such situations. Firth's test, despite being well calibrated and robust for testing low frequency and rare variants in unbalanced studies, lacks in computational efficiency as it also involves calculating the maximum likelihood under the full model. For instance, the projected computation time of the Firth's test to test 1500 phenotypes across 10 million SNPs is $\sim$ 117 CPU-years (2000 cases, 18000 controls). Thus, it is impractical to apply the Firth's test for analyzing large PheWAS datasets.

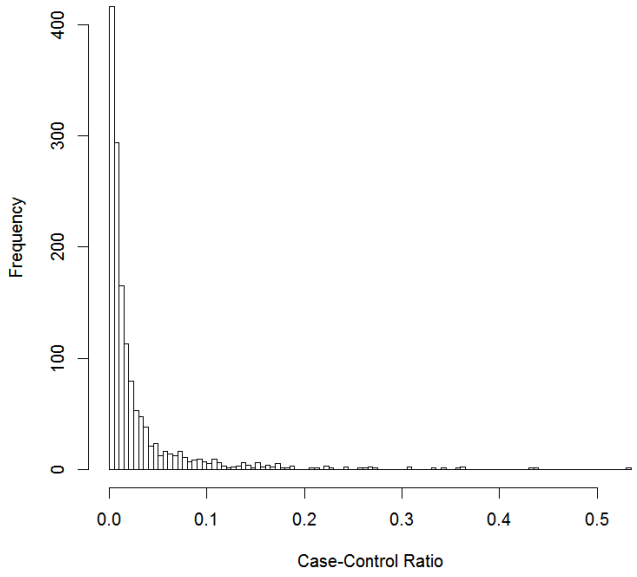


Figure 4.1: Histogram of case-control ratios of the 1448 phenotypes in the MGI data.

We propose a score-based single-variant test for binary phenotypes which is well calibrated for controlling the type I errors and can adjust for covariates even in extremely unbalanced case-control studies. Moreover, our test is computationally efficient and scalable to test thousands of phenotypes across millions of SNPs in large PheWAS datasets. Our proposed test (SPA) is based on the score statistics and

estimates the null distribution using the saddlepoint approximation (*Daniels*, 1954; *Barndorff-Nielsen*, 1990; *Kuonen*, 1999) instead of the normal approximation (*Feller*, 1945) traditionally used in score tests. We further develop an improvement of our test (fastSPA) which renders the most computationally challenging steps to be dependant only on the number of carriers (subjects with at least one minor allele) rather than the sample size. This improved test can substantially reduce the computation time, especially for low frequency and rare variants where the number of carriers is very low compared to the sample size. The projected computation time of our method to test for 1500 phenotypes across 10 million SNPs is $\sim$ 400 CPU-days (2000 cases, 18000 controls) which is more than a 100 times improvement over Firth's test. In addition, through the extensive simulation studies and analysis of the Michigan Genomics Initiative (MGI) data, we demonstrate that the proposed approach can control type I errors and is powerful enough to replicate known association signals.

4.2 Materials and Methods

4.2.1 Logistic Regression Model and Saddlepoint Approximation Method

We consider a case-control study with sample size n. For the i^{th} subject, let $Y_i = 1$ or 0 denote the case-control status, X_i the $k \times 1$ vector of non-genetic covariates including the intercept, and G_i the number of minor alleles ($G_i = 0, 1, 2$) of the variant to test. To relate genotypes to phenotypes, we use the following logistic regression model,

$$logit\left[Pr\left(Y_i = 1|X_i, G_i\right)\right] = X_i^T\beta + G_i\gamma \tag{4.1}$$

for $i = 1, 2, \ldots, n$ where β is a $k \times 1$ vector of coefficients of the covariates, and γ is the genotype log-odds ratio. Under this model, we are interested in testing for the genetic association by testing the null hypothesis $H_0 : \gamma = 0$. Let $\hat{\mu}_i$ be the estimate of $\mu_i = Pr\left(Y_i = 1|X_i\right)$, which is a probability to be a case under H_0. A score

statistic for γ from the model (4.1) is given by $S = \sum_{i=1}^{n} G_i (Y_i - \hat{\mu}_i)$. Suppose $X = \left(X_1^T, \ldots, X_n^T\right)$ is the $n \times k$ matrix of covariates, $G = (G_1, \ldots, G_n)^T$ is the genotype vector, W is a diagonal matrix with the i^{th} diagonal element being $\hat{\mu}_i (1 - \hat{\mu}_i)$, and $\tilde{G} = G - X \left(X^T W X\right)^{-1} X^T W G$ is a covariate adjusted genotype vector in which covariate effects are projected out from the genotypes (details given in Appendix D). Then S can be written as

$$S = \sum_{i=1}^{n} \tilde{G}_i (Y_i - \hat{\mu}_i), \tag{4.2}$$

and the mean and variance of S under H_0 are $E_{H_0}(S) = 0$ and $V_{H_0}(S) = \sum_{i=1}^{n} \tilde{G}_i^2 \hat{\mu}_i (1 - \hat{\mu}_i)$, respectively, where $\tilde{G}_i$ is the i^{th} element of $\tilde{G}$.

The traditional score test approximates the null distribution using a normal distribution which depends only on the mean and the variance of the score statistic. The p value can be obtained by comparing the observed test statistic, s and $N(0, V_{H_0}(S))$. Normal approximation works well near the mean of the distribution, but performs very poorly at the tails. The performance is especially poor when the underlying distribution is highly skewed, such as in unbalanced phenotypes (*Ma et al.*, 2013), since normal approximation cannot incorporate higher moments such as skewness. In addition, the convergence rate of normal approximation is $O\left(n^{-1/2}\right)$ (*Berry*, 1941; *Esseen*, 1942, 1956), which is not fast enough for rare variants.

Saddlepoint approximation was introduced by *Daniels* (1954) as an improvement over the normal approximation. Contrary to normal approximation, where only the first two cumulants (mean and variance) are used to approximate the underlying distribution, saddlepoint approximation uses the entire cumulant generating function. *Jensen* (1995) further showed that saddlepoint approximation has a relative error bound of $O\left(n^{-3/2}\right)$ making it a considerable improvement over the normal approximation.

To use saddlepoint approximation, we first derive the cumulant generating function (CGF) of S from the fact that $Y_i \sim Bernoulli(\mu_i)$ under H_0. Let $\hat{\mu}$ be an

$n \times 1$ vector with the i^{th} element being $\hat{\mu}_i$. From (4.2), the estimate of the cumulant generating function of the score statistic S is,

$$K(t) = log\left(E_{H_0}\left(e^{tS}\right)\right) = \sum_{i=1}^{n} log\left(1 - \hat{\mu}_i + \hat{\mu}_i e^{\tilde{G}_i t}\right) - t\sum_{i=1}^{n} \tilde{G}_i \hat{\mu}_i,$$

and the estimate of the first and second order derivatives of K are,

$$K'(t) = \sum_{i=1}^{n} \frac{\hat{\mu}_i \tilde{G}_i}{\left(1 - \hat{\mu}_i\right) e^{-\tilde{G}_i t} + \hat{\mu}_i} - \sum_{i=1}^{n} \tilde{G}_i \hat{\mu}_i, \quad K''(t) = \sum_{i=1}^{n} \frac{\left(1 - \hat{\mu}_i\right) \hat{\mu}_i \tilde{G}_i^2 e^{-\tilde{G}_i t}}{\left[\left(1 - \hat{\mu}_i\right) e^{-\tilde{G}_i t} + \hat{\mu}_i\right]^2}$$

respectively. We note that K, K' and K'' are plug-in estimates in which we plug in $\hat{\mu}_i$ instead of μ_i. Then, according to the saddlepoint method (*Barndorff-Nielsen*, 1990; *Kuonen*, 1999), the distribution of S at s can be approximated by

$$\Pr(S < s) \approx \tilde{F}(s) = \Phi\left\{w + \frac{1}{w}\log\left(\frac{v}{w}\right)\right\},$$

where $w = sgn\left(\hat{t}\right)\sqrt{2\left(\hat{t}s - K\left(\hat{t}\right)\right)}$, $v = \hat{t}\sqrt{K''\left(\hat{t}\right)}$, $\hat{t}$ is the solution to the equation $K'\left(\hat{t}\right) = s$, and Φ is the distribution function of a standard normal distribution.

4.2.2 Implementation Details and Approaches to Reduce the Computation Time

The saddlepoint approximation method involves finding the root of the saddlepoint equation $K'(t) = s$. It is easy to verify that K' is strictly increasing as $K''(t) > 0$ for all $-\infty < t < \infty$ and $s = \sum_{i=1}^{n} \tilde{G}_i\left(Y_i - \hat{\mu}_i\right)$ lies between $\lim_{t\to\infty} K'(t) = \sum_{i:\tilde{G}_i>0} \tilde{G}_i - \sum_{i=1}^{n} \tilde{G}_i \hat{\mu}_i$ and $\lim_{t\to-\infty} K'(t) = \sum_{i:\tilde{G}_i<0} \tilde{G}_i - \sum_{i=1}^{n} \tilde{G}_i \hat{\mu}_i$. Therefore a unique root exists, and we can use popular root-finding algorithms (Newton-Raphson (*Whittaker and Robinson*, 1967; *Press et al.*, 1992), bisection (*Press et al.*, 1992), secant (*Press et al.*, 1992), Brent's method (*Brent*, 1973)) to efficiently solve this equation. For our simulation studies and real-data applications we applied a combination of the

Newton-Raphson and bisection method to solve the saddlepoint equations.

The most computationally demanding step in this saddlepoint approximation method is calculating the cumulant generating function and its derivatives. Here we propose several approaches to reduce the computational complexities associated with these calculations.

4.2.2.1 Faster Calculation of the CGF Using a Partially Normal Approximation Approach

The most computationally intensive step in the saddlepoint method is the calculation of the cumulant generating function K and its derivatives. In each step of the root-finding algorithm we need to calculate K, K' and K'', each of which needs $O(n)$ computations. Using the fact that many elements of G are zeroes (i.e, homozygous major genotypes), we propose a fast computation method that speeds up the computation to $O(m)$, where m is the number of non-zero elements in G. Without loss of generality we assume that the first m subjects have at least one minor allele each and rests have homozygous major genotypes. We can then express S as $S = S_1 + S_2$ where $S_1 = \sum_{i=1}^{m} \tilde{G}_i \left(Y_i - \hat{\mu}_i\right)$ and $S_2 = \sum_{i=m+1}^{n} \tilde{G}_i \left(Y_i - \hat{\mu}_i\right)$. Let $Z = \left(X^T W X\right)^{-1} X^T W G$ and Z_l be the l^{th} element of Z. Then we can further express S_2 as,

$$
\begin{aligned}
S_2 &= \sum_{i=m+1}^{n} \tilde{G}_i \left(Y_i - \hat{\mu}_i\right) = \sum_{i=m+1}^{n} \left(0 - X_i Z\right) \left(Y_i - \hat{\mu}_i\right) \\
&= -\sum_{i=m+1}^{n} \sum_{l=1}^{k} X_{il} Z_l \left(Y_i - \hat{\mu}_i\right) = -\sum_{l=1}^{k} Z_l \sum_{i=m+1}^{n} X_{il} \left(Y_i - \hat{\mu}_i\right) \\
&= -\sum_{l=1}^{k} Z_l S_{2l}
\end{aligned}
$$

where $S_{2l} = \sum_{i=m+1}^{n} X_{il} \left(Y_i - \hat{\mu}_i\right)$. Now, if we assume that the non-genetic covariates are relatively balanced in the sample, then the normal distribution should be a good approximation for the null distribution of each S_{2l}. Since S_2 is a weighted sum of the S_{2l}s, we can also approximate the null distribution of S_2 using a nor-

mal distribution with mean and the variance under H_0 given by $E_{H_0}(S_2) = 0$ and $V_{H_0}(S_2) = \sum_{i=m+1}^{n} \tilde{G}_i^2 \hat{\mu}_i (1 - \hat{\mu}_i)$. Then, the CGF of S_2 can be approximated by,

$$K_2(t) = \frac{1}{2} t^2 V_{H_0}(S_2),$$

and the CGF of $S = S_1 + S_2$ can be approximated by,

$$K(t) = \sum_{i=1}^{m} \log\left(1 - \hat{\mu}_i + \hat{\mu}_i e^{\tilde{G}_i t}\right) - t \sum_{i=1}^{m} \tilde{G}_i \hat{\mu}_i + \frac{1}{2} t^2 V_{H_0}(S_2). \tag{4.3}$$

In order to calculate the first two terms at the right hand side of (4.3), we will need $\tilde{G}_i$s for $i = 1, \ldots, m$, which can be calculated in $O(m)$ computations since G only has m many non-zero elements and the quantity $X\left(X^T W X\right) X^T W$ can be pre-calculated. Then, the first two terms will require only $O(m)$ computations as both of them sums over m many elements. Next, the variance $V_{H_0}(S_2)$ can be further broken down into,

$$\begin{aligned}
V_{H_0}(S_2) &= \sum_{i=m+1}^{n} \tilde{G}_i^2 \hat{\mu}_i (1 - \hat{\mu}_i) = \sum_{i=m+1}^{n} (X_i Z)^2 \hat{\mu}_i (1 - \hat{\mu}_i) \\
&= \sum_{i=1}^{n} (X_i Z)^2 \hat{\mu}_i (1 - \hat{\mu}_i) - \sum_{i=1}^{m} (X_i Z)^2 \hat{\mu}_i (1 - \hat{\mu}_i) \\
&= Z^T \left(X^T W X\right) Z - \sum_{i=1}^{m} (X_i Z)^2 \hat{\mu}_i (1 - \hat{\mu}_i).
\end{aligned}$$

Since $X^T W X$ can be pre-calculated and Z is a $k \times 1$ vector, the first term requires $O(k)$ computations, and the second term requires $O(m)$ computations, which implies that the calculation of $V_{H_0}(S_2)$ requires $O(m)$ calculations assuming $k < m$, i.e, the number of non-genetic covariates is smaller than the number of subjects with at least one minor allele each. Hence, the cumulant generating function $K(t)$ can be calculated in $O(m)$ computations. Using similar arguments, we can further show that the derivatives $K'(t)$ and $K''(t)$ can also be calculated in $O(m)$ computations. Therefore, this partially normal approximation reduces the computational complexity of our test from $O(n)$ to $O(m)$, which is especially useful for rare variants, where m

is much smaller than n.

4.2.2.2 Using normal Approximation near the Mean for Faster Computation

Since the normal approximation behaves well near the mean of the distribution, we can use it to obtain the p value when the observed score statistic (s) lies close to the mean (zero). Moreover, saddlepoint approximation can be numerically unstable very close to the mean of the distribution. Such situations can also be avoided by using normal approximation near the mean. One possible approach is to use a fixed threshold in which we apply normal approximation to obtain the p value if the absolute value of the observed score statistic, $|s| < r\sigma$ where $\sigma = \sqrt{V_{H_0}(S)}$ and r is a pre-specified value. For example, we used $r = 2$ in our simulation studies and real data analyses. For a given level α, this approach does not inflate type I error rates if $r < \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the inverse function of the standard normal distribution function, $\Phi(x)$.

Alternatively, we can adaptively select the threshold using the error bound of the normal approximation given by the Berry-Esseen theorem. Suppose we are interested in controlling the type I error rate at level α. Let $F_n(x)$ be the true distribution function of the standardized score test statistic $S/\sqrt{V_{H_0}(S)}$. Then, according to Berry-Esseen theorem (*Berry*, 1941; *Esseen*, 1942, 1956), the maximum error bound in approximating $F_n(x)$ by $\Phi(x)$ is

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq B_n = C\left(\sigma^2\right)^{-3/2} \left(\sum_{i=1}^{n} \rho_i\right) \tag{4.4}$$

where $\rho_i = E_{H_0}\left[\left|\tilde{G}_i\left(Y_i - \hat{\mu}_i\right)\right|^3\right] = \tilde{G}_i^3 \hat{\mu}_i\left(1 - \hat{\mu}_i\right)\left[\hat{\mu}_i^2 + (1 - \hat{\mu}_i^2)\right]$, C is a constant. As of now, the best known estimate for C is 0.56, given by *Shevtsova* (2010). Suppose p_F and p_N are $F_n(x)$ and $\Phi(x)$ based p values. From the Berry-Esseen theorem, we

can show $p_N \leq p_F + B_n$. Suppose $q = B_n + \alpha/2$ and $r_\alpha = \Phi^{-1}(1-q)$. Then $p_N \geq q$ indicates $p_F \geq \alpha/2$. Therefore, we use $r_\alpha \sigma$ as a threshold at level α in which we will apply normal approximation if $|s| < r_\alpha \sigma$.

4.2.3 Numerical Simulations

To evaluate the computation times, type I error rates and power of the proposed method, we carried out extensive simulation studies. We considered three different case-control ratios: balanced with 10000 cases and 10000 controls, moderately unbalanced with 2000 cases and 18000 controls, and extremely unbalanced with 40 cases and 19960 controls. For each choice of case-control ratios, the phenotypes were simulated based on the following logistic model,

$$logit\left[\Pr\left(Y_i = 1\right)\right] = \beta_0 + X_{1i} + X_{2i} + \gamma G_i$$

where the two non-genetic covariates X_{1i} and X_{2i} were simulated from $X_{1i} \sim Bernoulli(0.5)$ and $X_{2i} \sim N(0,1)$. The intercept β_0 is chosen to correspond to prevalence 0.01. The genotype G_is were generated from a $Binomial(2,p)$ distribution where p was the MAF. The parameter γ represents the genotype log odds-ratio. To estimate computation times and type I error rates in realistic scenarios, the MAF (p) was randomly sampled from the MAF distribution in the MGI data. For the computation time comparisons, we simulated 10^4 variants with $\gamma = 0$. For the type I error comparisons, we simulated 10^9 variants with $\gamma = 0$ and recorded the number of rejections at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}. We also used fixed MAFs to evaluate the effect of MAFs to computation time and type I error rates. For the power calculations, we considered two different choices for MAF, $p = 0.01$ and 0.05, and wide ranges of γ (Figure 4.5). For each choice of p and γ we generated 5000 variants.

We compared the computation times of seven different tests: traditional score test

using normal approximation (Score); the saddlepoint approximation based test with the standard deviation threshold at 0.1 and 2 (SPA-0.1 and SPA-2); the fast saddlepoint approximation based test with the partially normal approximation improvement and the standard deviation threshold at 0.1 and 2 (fastSPA-0.1 and fastSPA-2); the fastSPA test with the Berry-Esseen bound threshold at level (fastSPA-BE); and the Firth's penalized likelihood test (Firth). Next, we compared the empirical type I errors and power curves for fastSPA-2, score and Firth tests at level 5×10^{-8}. Since performing the Firth test 10^9 times, which is required to estimate type I error rates at level 5×10^{-8}, is practically impossible due to the heavy computational burden of the Firth test, we performed a hybrid approach in which we used the Firth test only when the fastSPA-2 p values were smaller than 5×10^{-3}. For the power comparison, since the score test has extremely inflated type I errors in the unbalanced and extremely unbalanced case-control scenarios (as shown in Section 4.3), it may not be appropriate to directly compare the power of the score test to the other two tests at the same nominal α level. In order to provide a more meaningful comparison, we compared their powers at their empirical α levels where their empirical type I errors become 5×10^{-8}. The empirical α levels were selected based on the type I error simulations with variants simulated with MAF randomly sampled from the MAF distribution of the MGI data. This approach is similar to performing resampling (e.g., permutation) to control family-wise error rates. We also estimated the powers at the nominal fixed $\alpha = 5 \times 10^{-8}$. In order to compare the p values resulted from different tests, we also simulated 5×10^6 variants with MAFs randomly sampled from the MAF distribution of the MGI data. We further compared the inflation factors of the genomic controls at different p value quantiles for fastSPA-2, fastSPA-BE and fastSPA-0.1 in order to explore the effect of the standard deviation threshold on the inflation factor.

4.2.4 Michigan Genomics Initiative (MGI) Data Application

To illustrate the performance of the proposed methods in real data application, we analyzed four selected phenotypes in the MGI data. The main goal of MGI is to create an institutional repository of genetic data together with rich clinical phenotypes for a broad portfolio of future medical research. DNA from blood samples of $> 20,000$ surgical patients at the University of Michigan Health System was genotyped (with their informed consent) on the Illumina HumanCoreExome v12.1 array, which is a combination GWAS plus exome array comprised of $> 500,000$ single nucleotide polymorphisms. Genotypes of the Haplotype Reference Consortium (HRC) (*McCarthy et al.*, 2016) (chromosome 1-22: HRC release 1; chromosome X: HRC release 1.1) were imputed into the phased MGI genotypes (SHAPEIT2 (*Delaneau et al.*, 2013) on autosomal chromosomes and Eagle2 (*Loh et al.*, 2016) on chromosome X) using Minimac3 (*Das et al.*, 2016). Excluding variants with low imputation quality ($R^2 < 0.3$) resulted in dense mapping at over 39 million quality-imputed genetic markers.

Phenotypes derived from 8,940 ICD-9 billing codes were classified into 1,815 PheWAS disease states of shared disease etiology, of which 1,448 had at least 20 cases. Standard code translations were used to convert the taxonomy of diagnostic ICD-9 codes into PheWAS code groups (PheWAS code translation table version 1.2 (*Carroll et al.*, 2014)). Cases were derived from electronic health records for individuals with at least 2 encounters with an ICD-9 billing code. This is a typical example of many large-scale PheWASs that are being conducted in recent days. In order to compare our proposed fastSPA-2 test with the traditional score test and the current gold standard Firth test in analyzing such PheWAS data, we performed genome-wide association analyses for 4 selected traits, Skin Cancer (PheWAS code: 172), Type-2 diabetes (PheWAS code: 250.2, [MIM: 125853]), Primary Hypercoagulable state (PheWAS code: 286.81, [MIM: 188055]) and Cystic Fibrosis (PheWAS code: 499, [MIM: 219700]), in 18,267 unrelated individuals of European ancestry, with ad-

justment for age, sex, and four principal components. Genotyped samples with any missing covariate information were excluded from analysis. Since imputation quality is low for very rare variants (*McCarthy et al.*, 2016), we excluded the imputed variants with MAF < 0.001 in our main analysis, which resulted in 13 million variants. For the Firth test, we used the hybrid approach used in the type I error simulation in which the Firth test was performed only when the fastSPA-2 p value was smaller than 5×10^{-3}.

4.3 Results

4.3.1 Numerical Simulations

We examine the computation time, type I error control and power of the proposed fastSPA and two existing approaches, score and Firth tests, across ranges of case-control imbalance and MAFs.

4.3.1.1 Comparison of Computation Times

The projected computation times for testing 1500 phenotypes across 10 million variants using different testing methods are presented in Figure 4.2. To obtain computation time under realistic scenarios of the MAF distribution, the MAFs of the simulated SNPs were randomly sampled from the MAF spectrum of the MGI data (Figure E.1). The fastSPA-2 test performs 100-300 times faster than the Firth's test. In the unbalanced case-control setup of 2000 cases and 18000 controls, for example, the Firth's test takes 117 CPU-years whereas fastSPA-2 only takes 1.09 CPU-years to analyze 10 million SNPs across 1500 phenotypes. This indicates that on a cluster with 100 CPU cores, the proposed test would require 4 days (without data reading) but the Firth' test would need more than a year. When we compare fastSPA and SPA, fastSPA-0.1 performs 4-6 times faster than SPA-0.1 (e.g. 2.90 vs 12.32

CPU years when case:control = 2000:18000), and fastSPA-2 performs 1.5-2 times faster than SPA-2 (e.g. 1.09 vs 1.62 CPU years when case:control = 2000:18000). Expectedly, the computation time for fastSPA-BE is in between the computation times for fastSPA-2 and fastSPA-0.1. fastSPA-BE performs 1.3-1.8 times faster than fastSPA-0.1 and 1.6-2.8 times slower than fastSPA-2 (eg. 1.09, 1.86, 2.9 CPU years for fastSPA-2, fastSPA-BE and fastSPA-0.1 when case:control = 2000:18000).

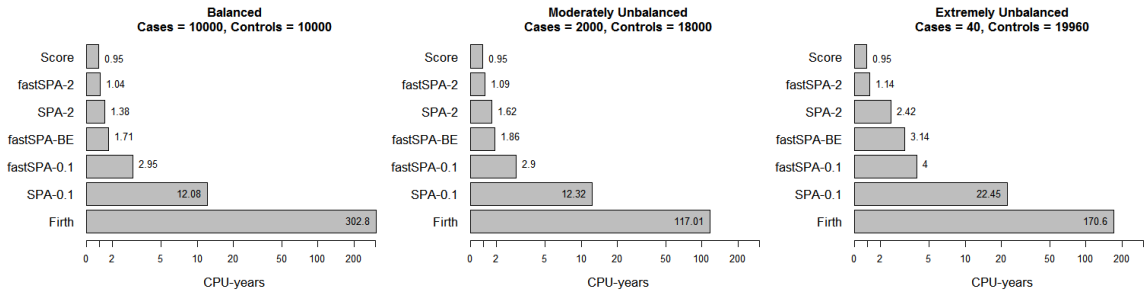


Figure 4.2: Projected computation times for testing 10 million variants across 1500 phenotypes using different single-variant tests with MAFs sampled from the MAF distribution of the MGI data. The computation times are based on testing 10000 simulated variants on an Intel i7 2.70GHz processor, and then projecting it onto a PheWAS study with 10 million variants and 1500 phenotypes.

We also recorded the computation times for variants with three different fixed MAFs 0.1, 0.01 and 0.001 in order to assess the effect of MAF on the performance of the tests. Similar to Figure 4.2, Table 4.1 also shows the superior performance of fastSPA-2 compared to all other tests. Moreover, while the computation time of SPA increases with decreasing MAFs, which may be due to the slow convergence caused by the discrete nature of the underlying distribution, fastSPA requires less computation time for rarer variants (smaller MAFs) compared to more common variants (larger MAFs). This demonstrates the potential of the partially normal approximation improvement in terms of faster computation of the p values, especially for low-frequency and rare variants.

Case:Control	MAF	Score	SPA-0.1	fastSPA-0.1	fastSPA-BE	SPA-2	fastSPA-2	Firth
10000:10000	0.1	20	214	75	37	28	23	7251
	0.01	19	225	38	35	27	20	6918
	0.001	19	242	33	36	30	20	5304
2000:18000	0.1	21	256	84	37	36	24	3940
	0.01	20	284	39	36	35	21	4312
	0.001	19	326	34	41	40	20	3804
40:19960	0.1	21	376	98	70	38	24	3615
	0.01	20	477	42	58	44	21	3598
	0.001	20	647	38	51	79	21	3525

Table 4.1: Computation times for various tests when testing 10000 simulated variants with different MAFs. All computation times are in CPU-seconds on an Intel i7 2.70GHz processor.

4.3.1.2 Type I Error Comparison

The type I error rates from 10^9 simulated datasets are presented in Figure 4.3. Due to the heavy computation burden for testing these extremely large numbers of datasets, in this comparison, we only considered the traditional score test, fastSPA-2, and the hybrid version of the Firth test, in which we used the Firth test only when the fastSPA-2 p values were smaller than 5×10^{-3}. We note that both fastSPA-2 and Firth tests had well calibrated QQ plots up to 10^{-6} p values (Figure 4.6), and whenever fastSPA-2 p values $> 5 \times 10^{-3}$, Firth test p values $> 4.8 \times 10^{-4}$ (see Section 4.3.1.4), indicating that the hybrid approach can provide very accurate estimate of the type I error rates of the Firth test at very stringent α levels.

The traditional score test had greatly inflated type I error rates for moderately unbalanced and extremely unbalanced case-control ratios, whereas fastSPA-2 can control the type I error in such situations. At the genome-wide significance level of $\alpha = 5 \times 10^{-8}$, for example, the empirical type I error rates of the score test were 32 (1.63×10^{-6}, when case:control = 2000:18000) and 26600 (1.33×10^{-3}, when case:control = 40:19960) times higher than the nominal $\alpha = 5 \times 10^{-8}$. In contrast, the fastSPA-2 had empirical type I error rates nearly identical (4.9×10^{-8}, when

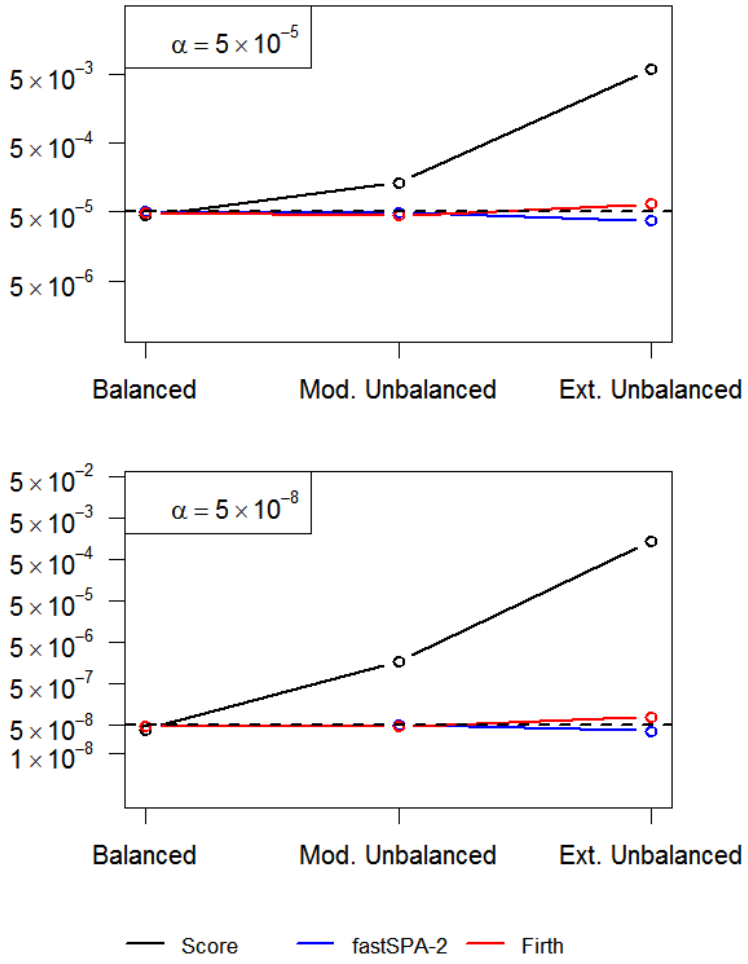


Figure 4.3: Type I error comparison between the traditional score test, fastSPA-2 and Firth tests for variants simulated with MAFs sampled from the MAF distribution of the MGI data. Type I error rates were estimated based on 10^9 simulated datasets. From left to right on the x-axis, the plots consider case:control = 10000:10000 (Balanced), 2000:18000 (Moderately Unbalanced) and 40:19960 (Extremely Unbalanced), respectively. The top and the bottom panels show empirical type I error rates at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8} levels respectively.

case:control = 2000:18000) or slightly lower (3.5×10^{-8}, when case:control = 40:19960) than the nominal 5×10^{-8}. The Firth test also had well controlled type I error rates in the balanced and moderately unbalanced case-control scenarios (4.7×10^{-8} and 4.9×10^{-8}, respectively at $\alpha = 5 \times 10^{-8}$). Interestingly, it shows slight inflation (7.8×10^{-8} at $\alpha = 5\times10^{-8}$) in the extremely unbalanced scenario. We also estimated empirical type I error rates at six different MAFs (Figure 4.4). The score test had deflated type I error rates for low-frequency and rare variants for the balanced case-control ratio and inflated and extremely inflated type I error rates for moderately and severely unbalanced case-control ratios. The fastSPA-2 method had overall well controlled type I error rates regardless of MAFs and case-control ratios. The Firth test had either well controlled or slightly conservative type I error rates when the case-control ratio was balanced or moderately unbalanced. However, when the case control ratio was extremely unbalanced, the Firth test had inflated type I error rates especially when the minor allele count was small (eg. 1.33×10^{-7} and 1.47×10^{-7} for MAF = 0.0005 and 0.001 respectively at $\alpha = 5 \times 10^{-8}$ when case:control=40:19960).

4.3.1.3 Power Comparison

Next, we compared the power curves of fastSPA-2, score and Firth tests. Note that the Firth test (*Firth*, 1993) is a current gold standard method. Since the traditional score test had greatly inflated type I error rates, we compared the empirical powers of different tests at their test-specific empirical α levels. Figure 4.5 shows power by odds ratios when the MAF of the variant was 0.05 (top panel) and 0.01 (bottom panel). As expected, the power is higher when the case-control ratio is balanced. The empirical powers of fastSPA-2 and the Firth test were nearly identical for all case-control ratios and MAFs, which suggests that our proposed test does not suffer from any loss in power compared to the Firth test. The empirical powers of the score test were almost

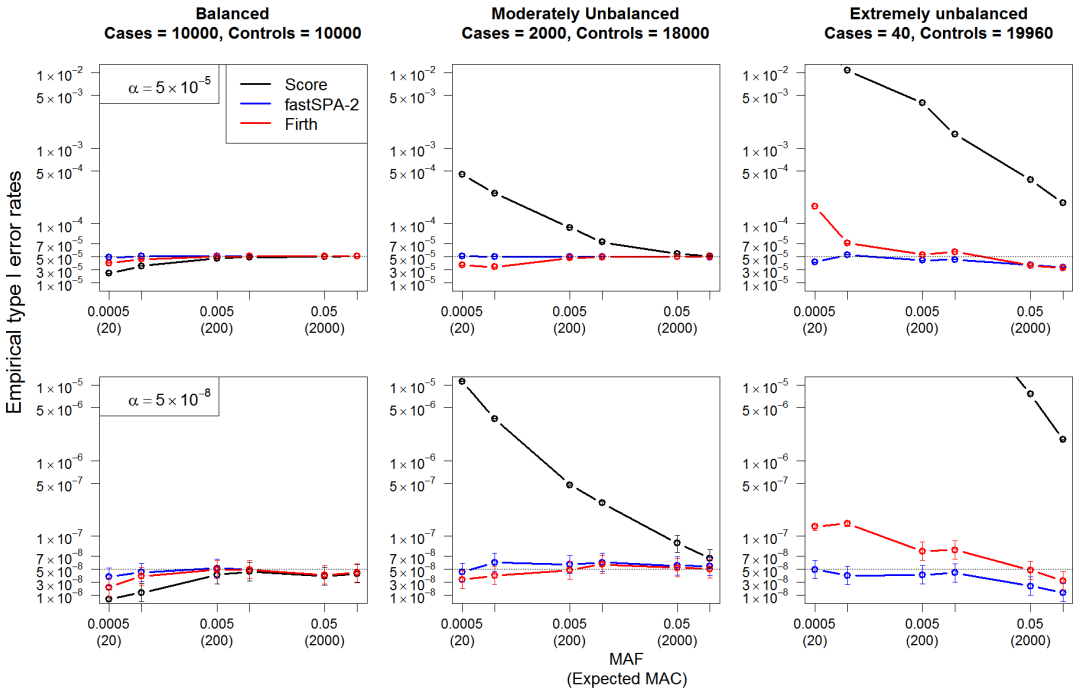


Figure 4.4: Type I error comparison at different MAFs between the traditional score test, fastSPA-2 and Firth tests. The top and bottom panels show empirical type I error rates at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. From left to right, the plots consider case:control = 10000:10000, 2000:18000 and 40:19960, respectively. In each plot x-axis represents MAF with expected MAC in the parenthesis, and y-axis represents empirical type I error rates. Empirical type I error rates were estimated based on 10^9 simulated datasets. 95% confidence intervals at different MAFs are also presented.

identical to those of fastSPA-2 and Firth test for the balanced case-control ratio. However, the score test showed substantially lower power than the other two tests for the unbalanced case-control ratios due to the very small empirical α levels, and the power gap is especially large when the case-control ratio is extremely unbalanced. The simulation results clearly show that the proposed approach improves power over the score test when type I error rates were properly controlled. When we used nominal $\alpha = 5 \times 10^{-8}$ level instead of the empirical α levels, score test had higher power than the other two approaches as expected (Figure E.2), since its type I error rates were not controlled.

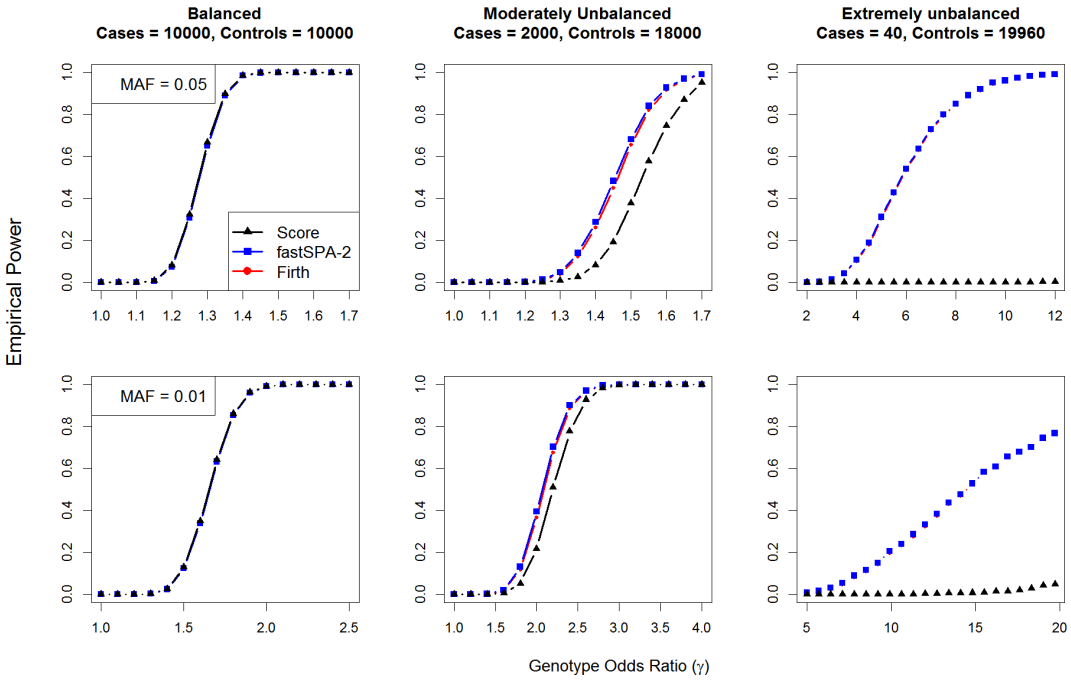


Figure 4.5: Empirical power curves for the traditional score, fastSPA-2 and Firth tests at their empirical α levels where their empirical type I errors are equal to 5×10^{-8}. Top panel considers MAF = 0.05 and bottom panel considers MAF = 0.01. From left to right, the plots consider case:control = 10000:10000, 2000:18000 and 40:19960, respectively. In each plot x-axis represents genotype odds ratios and y-axis represents the empirical power. Empirical power was estimated from 5000 simulated datasets.

4.3.1.4 P Value and Inflation Factor (λ) Comparison

To compare p value distributions of various tests, we generated QQ plots and calculated the inflation factor (λ) of the genomic control. Figure 4.6 suggests strong deflation (smaller than expected) in the p values based on the traditional score test in the moderately unbalanced and extremely unbalanced case-control setups, whereas fastSPA-2, SPA-2 and Firth tests resulted in well calibrated QQ plots, which suggest that these methods can control for type I errors. Moreover, the minimum Firth p value was 4.8×10^{-4} for the variants with fastSPA-2 p value $> 5 \times 10^{-3}$ among all case-control setups, which justifies our hybrid approach of performing Firth test only when fastSPA-2 p value $< 5 \times 10^{-3}$ in the type I error simulation studies.

None of fastSPA-2, fastSPA-BE and fastSPA-0.1 tests showed any inflation or deflation in genomic controls (λ) in the balanced and moderately unbalanced case-control setups (Table E.1). In the extremely unbalanced case-control setup, fastSPA-2 resulted in greatly deflated inflation factor ($\lambda = 0.48$) at the median of p value ($q = 0.5$). Interestingly fastSPA-BE and fastSPA-0.1 resulted in inflated λ (both having $\lambda = 1.83$) at $q = 0.5$, which may be due to the discrete nature of p values. When λ was measured at p value quantiles $q = 0.01$ and 0.001, however, all three tests provided λ very close to unity.

4.3.2 MGI Data Analysis

We applied the traditional score test, Firth test and the fastSPA-2 method to the MGI data with four phenotypes, Skin Cancer, Type-2 diabetes, Primary Hypercoagulable state, and Cystic Fibrosis, which were selected based on case-control ratios. Skin Cancer (2359 cases, 15265 controls) and Type-2 diabetes (1987 cases, 14906 controls) were moderately unbalanced, whereas Primary Hypercoagulable state (168 cases, 16401 controls) and Cystic Fibrosis (28 cases, 18212 controls) were extremely unbalanced phenotypes.

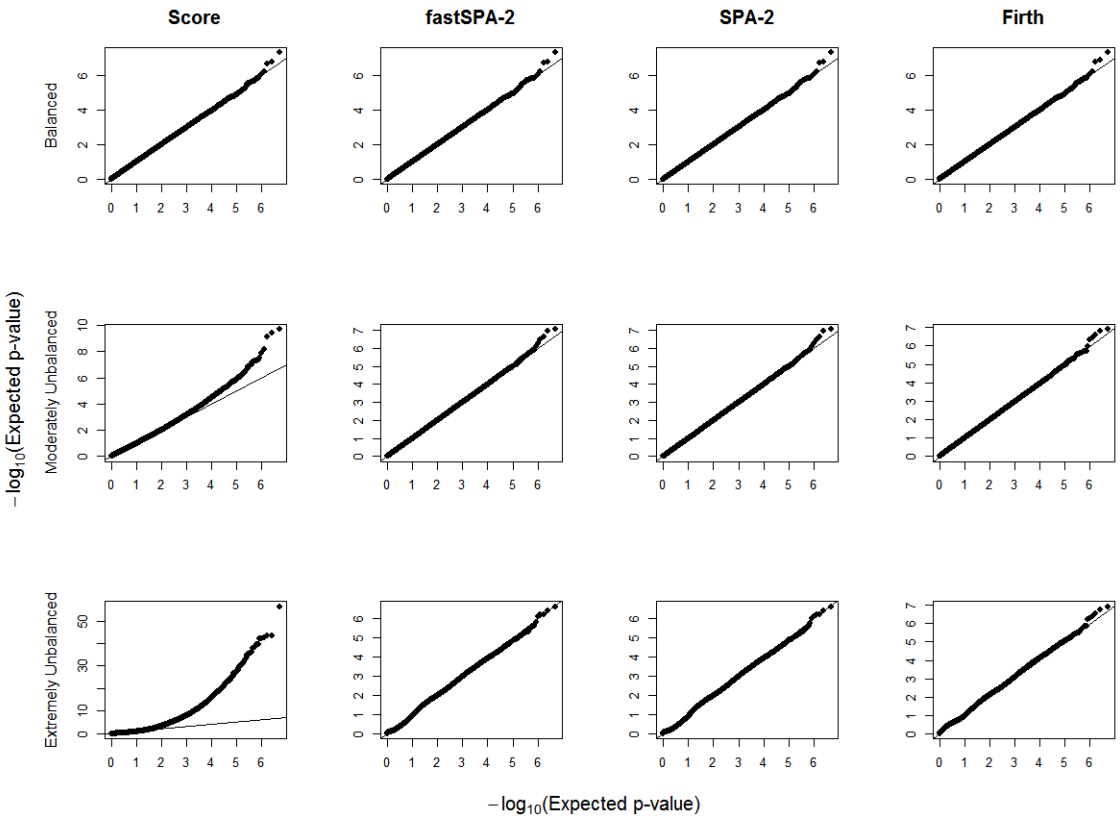


Figure 4.6: QQ plots for the traditional score, fastSPA-2, SPA-2 and Firth tests on 5×10^6 simulated variants with MAF randomly sampled from the MAF distribution of the MGI data. The top, middle and bottom panels show QQ plots in the balanced (case:control = 10000:10000), moderately unbalanced (case:control = 2000:18000) and extremely unbalanced (case:control = 40:19960) case-control scenarios respectively. In each plot, x-axis represents $-log_{10}$ expected p values, and y-axis represents $-log_{10}$ observed p values.

The Manhattan plots (Figure 4.7) show that the traditional score test produced a large number of potentially spurious associations for all of these phenotypes, whereas all of the significant variants from our proposed test at the genome-wide significant level of $\alpha = 5 \times 10^{-8}$ can be verified as truly associated with the phenotypes based on previous findings (Table 4.2). In the analysis of Skin Cancer, variants in or near *IRF4* (MIM: 601900), *MC1R* (MIM: 155555), *RALY* (MIM: 614663) and *SLC45A2* (MIM: 606202) were significant at and all of these four genes were previously identified (*Zhang et al.*, 2013; *Sulem et al.*, 2007; *Jacobs et al.*, 2015; *Liu et al.*, 2015; *Barrett et al.*, 2011; *Nan et al.*, 2009) as associated with pigmentation traits and skin cancers. In the other traits, variants in *TCF7L2* (MIM: 602228), *F5* (MIM: 612309) and *CFTR* (MIM: 602421) were significantly associated with Type2 diabetes (*Scott et al.*, 2006), Primary Hypercoagulable State *Bertina et al.* (1994) and Cystic Fibrosis (*Kerem et al.*, 1989), respectively, and all of these genes are well known to be associated with the risk of each disease. The QQ plots (Figure 4.8) also suggest that the p values based on the traditional score test are much smaller than expected, especially for low-frequency and rare variants, whereas the p values based on fastSPA-2 closely follow the uniform distribution. We also observed the Manhattan plots (Figure E.3) including the imputed variants with MAF < 0.001 in the analysis. The inclusion of rarer variants resulted in extreme inflation in the number of potentially spurious associations for the traditional score test. However, our proposed test still produced none to very few new associations. The Manhattan plots and QQ plots for the Firth test were almost identical to those of our proposed test.

Further, based on the p values from our proposed test, we obtained the inflation factor λ of the genomic control at different p value quantiles (q) and different MAF cut-offs (Table E.2). Only the imputed variants were removed when we used different MAF cutoffs. The SNPs present on the Illumina HumanCoreExome v12.1 array were preserved. To evaluate whether using a smaller standard deviation threshold (r)

Phenotype	Location	dbSNP ID	Nearest Gene	Alleles	MAF	p value	Previous Findings
Skin Cancer	6:396321	rs12203592	*IRF4*	$C > T$	0.16	6.71×10^{-18}	*Zhang et al.* (2013); *Sulem et al.* (2007)
	16:89986117	rs1805007	*MC1R*	$C > T$	0.077	1.86×10^{-14}	*Jacobs et al.* (2015); *Liu et al.* (2015)
	20:32538391	rs62211989	*RALY*	$G > C$	0.075	5.59×10^{-13}	
	5:33951693	rs16891982	*SLC45A2*	$C > G$	0.038	7×10^{-9}	*Liu et al.* (2015); *Barrett et al.* (2011) *Nan et al.* (2009)
Type-2 Diabetes	10:114754071	rs34872471	*TCF7L2*	$T > C$	0.29	3.4×10^{-11}	*Scott et al.* (2006)
Primary Hypercoagulable State	1:169519049	rs6025	*F5*	$T > C$	0.029	4.9×10^{-39}	*Bertina et al.* (1994)
Cystic Fibrosis	7:117299434	rs113827944	*CFTR*	$G > A$	0.018	3.11×10^{-15}	*Kerem et al.* (1989)

Table 4.2: Significant SNP-phenotype associations based on fastSPA-2 test on MGI data and previous findings confirming such associations.

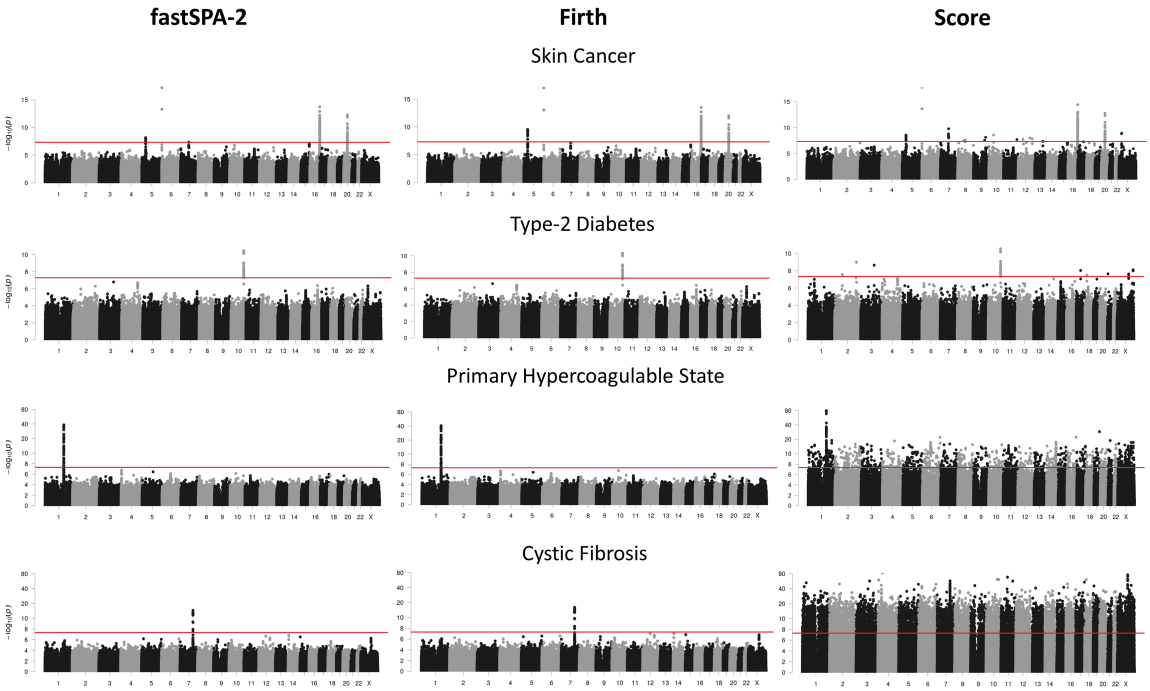


Figure 4.7: Manhattan plots for four different phenotypes from the MGI data (excluding imputed variants with MAF ≤ 0.001). From left to right, the three panels show associations based on the fastSPA-2, Firth, and traditional score tests, respectively. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

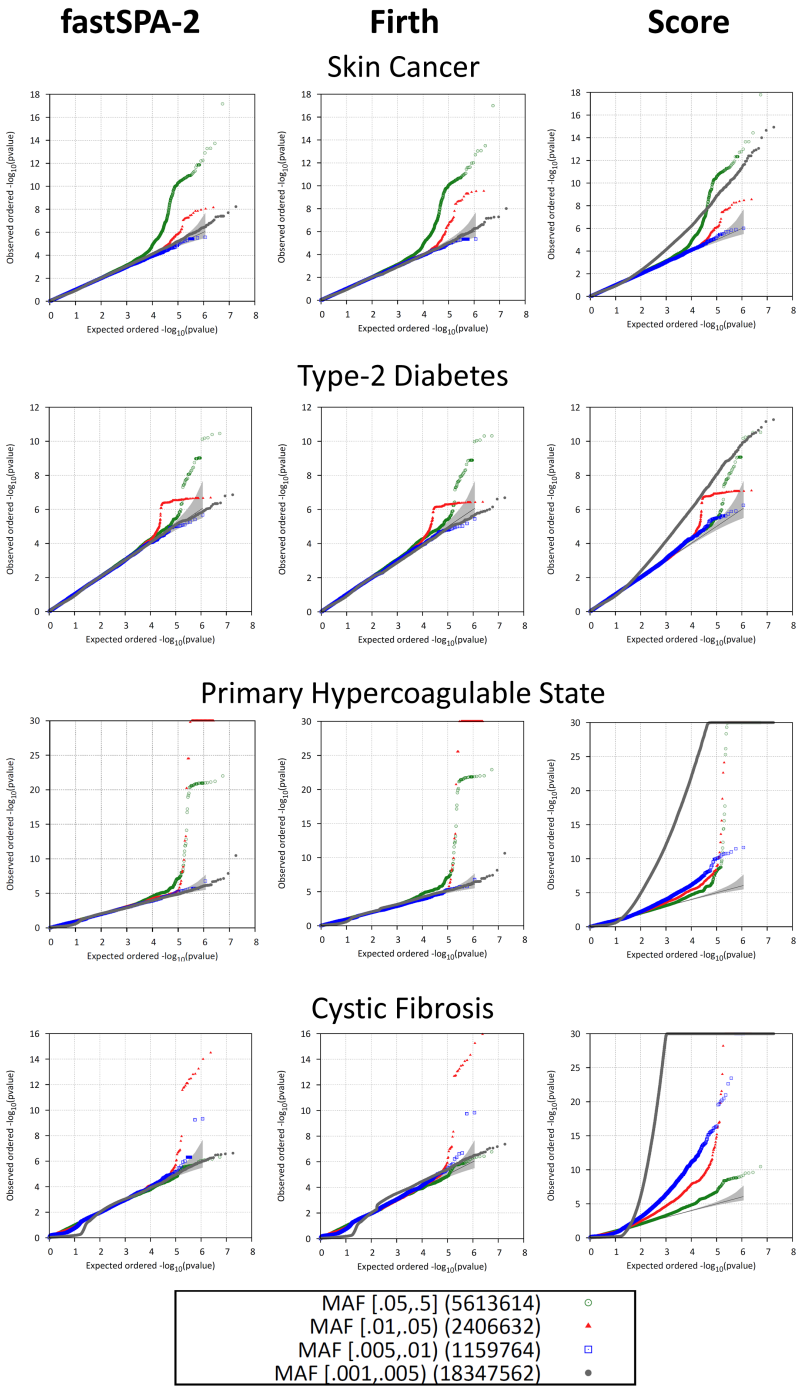


Figure 4.8: QQ plots for four different phenotypes from the MGI data. From left to right, the three panels show the QQ plots based on the fastSPA-2, Firth, and traditional score tests, respectively. The plots are color-coded based on different MAF categories.

improves the estimation of λ, we also applied fastSPA with $r = 0.1$ (i.e fastSPA-0.1), and fastSPA with the Berry-Esseen bound threshold at level (fastSPA-BE) on these four phenotypes. When all the variants were included in the analysis, there was slight inflation ($\lambda = 1.11$, type 2 diabetes) or great deflation ($\lambda = 0.12$, Cystic fibrosis) at the median level for fastSPA-2. However, the genomic controls are very close to unity at $q = 0.01$ and $q = 0.001$. If we only consider the variants with MAF > 0.001, then fastSPA-2 does not show any significant inflation in λ at the median for Skin Cancer, Type-2 Diabetes, and Primary Hypercoagulable State. Although it shows a deflated genomic control for Cystic Fibrosis ($\lambda = 0.63$) due to the discrete nature of the underlying distribution. However, if we exclude the rare variants and consider only the variants with MAF > 0.01, then all four of the phenotypes show λ very close to unity. Both fastSPA-0.1 and fastSPA-BE show no significant inflation or deflation in λ at all quantiles and MAF cut-offs, except for Cystic Fibrosis (both having $\lambda = 1.27$) when all the variants are considered and genomic control is measured at the median level.

4.4 Discussion

In this chapter, we proposed a fast and scalable test to analyze large PheWAS datasets which is well calibrated even in extremely unbalanced case-control settings. The method uses computationally efficient saddle point approximation to accurately calculate p values of score test statistics. We further proposed an improved version of our test which substantially reduces the computation time, especially for low-frequency and rare variants. Our proposed test can also adjust for additional covariates. Through extensive numerical studies we demonstrated that our test can perform 100 – 300 times faster than the currently used Firth's test while retaining similar power and well controlled type I error rates. MGI data analysis illustrates that by applying the proposed method to PheWAS, we can identify true association

signals while controlling for type I error, even for traits with a very small number of cases and a large number of controls.

Our test calculates p values based on the traditional score test if the score statistics lie sufficiently close to the mean. Even though normal approximation is accurate near the mean, those p values may not be well calibrated. In such cases, since the median p values might come from the traditional score test, we can encounter slightly inflated or deflated inflation factor at median. When the case control ratio is extremely unbalanced, this phenomenon is more pronounced. One way to circumvent this issue is to measure the inflation factor at more extreme quantiles (0.01, 0.001 etc.), or to exclude rare variants when estimating the inflation factor. Another approach is to decrease the standard deviation threshold so that the median p values come from the saddlepoint approximation. In the MGI data analysis, fastSPA-0.1 produced substantially improved inflation factor estimates than fastSPA-2. However, the use of threshold 0.1 instead of 2 would increase the computation time $\sim 3-4$ times. The Berry-Esseen threshold can be viewed as a compromise between these two thresholds. If there is no restriction in computational resource, we recommend to use fastSPA-0.1 so that most of the p values are calculated using the saddlepoint approximation. If computational resource is limited, or researchers want to obtain results quickly, either a larger threshold (i.e fastSPA-2) or Berry-Esseen bound can be a better choice.

As sequencing costs continue to drop, whole-exome or whole-genome sequencing will be used for PheWAS to identify rare variants associated with clinical phenotypes (*Collins and Varmus*, 2015). In rare variant association analysis, gene or region based multiple variant tests (*Lee et al.*, 2014a) are commonly used to improve power. When case-control ratios are unbalanced, popular rare variant tests, including burden tests, SKAT and SKAT-O, can also have substantially inflated type I error rates. Although resampling based approaches (*Lee et al.*, 2015) have been developed to address this problem, the existing methods are not fast enough to be used in PheWAS. One

possible approach is first to adjust single-variant score statistics using SPA and then to use the adjusted score statistics to control for the type I error. We left it for future research.

In summary, we have proposed an accurate and scalable method for PheWAS data analysis. With the growing effort to build large research cohorts for precision medicine (*Collins and Varmus*, 2015), future PheWAS would have hundreds of thousands of samples and hundreds of millions of variants. Our method will provide a scalable solution for this large-scale problem and contribute to finding genetic component of complex traits. All our tests are implemented in the R package **SPAtest**.

CHAPTER V

Robust Meta-Analysis of Biobank-based Genome-wide Association Studies with Unbalanced Binary Phenotypes

5.1 Introduction

Genome-wide scale phenome-wide association analysis (*Hebbring*, 2014) is gaining increasing attention in the human genetics community in the recent years. The availability of detailed phenotypic information from the electronic health record (EHR) systems in large biobanks as well as the recent advancements in genotyping and imputation technologies (*Das et al.*, 2016; *Marchini and Howie*, 2010) are allowing researchers to phenotype thousands of traits and genotype tens of millions of variants in large cohort studies. Several biobank studies, including UK Biobank (*Bycroft et al.*, 2017), Michigan Genomics Initiative (`https://www.michigangenomics.org/`) and Nord-Trøndelag Health Study (*Krokstad et al.*, 2013) currently attempt to test for associations in all genotype-phenotype pairs, which results in billions of tests. These large-scale analyses have great potential to find novel genotype-phenotype associations, which will help uncover underlying molecular mechanism of clinical phenotypes.

In a typical phenome-wide association study (PheWAS) in biobanks, most of

the phenotypes are binary with unbalanced (1 : 5) or often extremely unbalanced (1 : 500) case-control ratios, which results in performing 1000s of unbalanced case-control GWASs. For example, $\sim$ 1400 case-control studies in the UK Biobank interim release data have more than 100 controls per case (see histogram in Figure 5.1). Under such case-control imbalance, the standard asymptotic tests such as the Wald test, score test and likelihood ratio test can severely inflate the type I errors resulting in several spurious associations, especially for the low frequency ($0.01 < \text{MAF} < 0.05$) and rare ($\text{MAF} < 0.01$) variants (*Dey et al.*, 2017; *Ma et al.*, 2013). To obtain well-calibrated p values in such situations, *Ma et al.* (2013) proposed to use the Firth's penalized likelihood ratio test (*Firth*, 1993). Since the Firth's test is computationally too expensive to be used for billions of association tests, we developed a fast saddle-point approximation-based score test, fastSPA (Chapter IV,*Dey et al.* (2017)), which is computationally much faster than the Firth's test.

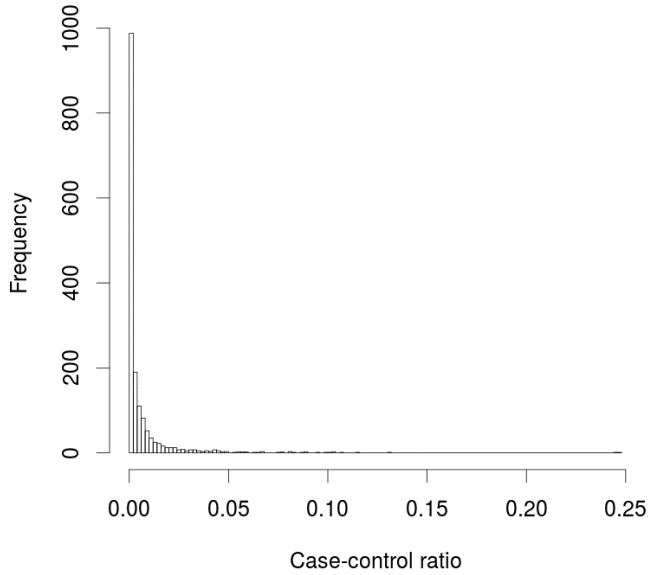


Figure 5.1: Histogram of case-control ratios of the 1688 binary phenotypes in the UK Biobank interim release data.

As more and more association results from different biobanks become available,

meta-analyzing (*Evangelou and Ioannidis*, 2013) the results from the unbalanced GWASs is the logical next step to improve the power to detect novel genotype-phenotype associations. Z-score-based approach, (*Cooper et al.*, 2009) which converts p values to normal Z-scores for combining multiple study p values, has been a standard meta-analysis method in GWASs (*Evangelou and Ioannidis*, 2013). However, even though p values from fastSPA and Firth's test are well calibrated in a single study, combining them through Z-score method can fail to control for type I errors. *Ma et al.* (2013) has shown that combining Firth's test-based p values through Z-score method can produce conservative or anti-conservative behaviors especially when the case-control ratio is unbalanced and the variant minor allele count (MAC) is small. This may be because the study-specific p values have discrete distribution due to case-control imbalance and small MAC. As shown in our simulation studies, the same problem also occurs in the meta-analysis using fastSPA-based p values. To facilitate the meta-analysis of the biobank-based GWASs, we need a robust method to control for type I errors regardless case-control ratios and MAC.

In this chapter, we first evaluate the performance of the Z-score-based meta-analysis procedure using the fastSPA test-based p values under extensive simulation settings and real datasets, and propose two alternative meta-analysis strategies to obtain well-calibrated meta-analysis p values. The first method involves sharing the observed within-study score statistics and the cumulant generating functions (CGF) of those score statistics using a spline-based approach, which will be used to carry out saddlepoint approximation to obtain the meta-analysis p value. The second method involves sharing the overall number of homozygous minor and heterozygous genotypes for each genetic variant, in addition to the case-control sample size and p value shared in the Z-score-based meta-analysis strategy. The additional information facilitates approximating the distributions of the study-specific score statistics, which can be discrete, asymmetric and different from the traditionally used normal distribution.

Through extensive simulation studies and an analysis of the UK Biobank data, we show that the proposed methods can control the type I error rates and retain similar power as a joint analysis as well as being scalable to large-scale PheWASs.

5.2 Methods

5.2.1 Model for Single Study Association Test and Saddlepoint Approximation (SPA)

We consider J case-control studies, where the j^{th} study has sample size n_j. Within each individual study, we follow the regression model and testing procedure described in (*Dey et al.*, 2017). For the i^{th} subject in the j^{th} study, let $Y_i^{(j)} = 1$ or 0 denote the case-control status, $X_i^{(j)}$ denote the $k \times 1$ vector of non-genetic covariates (including the intercept) and $G_i^{(j)} = 0, 1, 2$ denote the number of minor alleles of the variant to be tested. Let $\beta^{(j)}$ be the $k \times 1$ vector of coefficients for the non-genetic covariates and $\gamma^{(j)}$ be the genotype log odds ratio. We use the following logistic regression model to perform association test in the j^{th} study.

$$\text{logit}\left[\Pr\left(\mathrm{Y_i^{(j)} = 1 | X_i^{(j)}, G_i^{(j)}}\right)\right] = \mathrm{X_i^{(j)T}}\beta^{(j)} + \mathrm{G_i^{(j)}}\gamma^{(j)} \tag{5.1}$$

Let $\hat{mu}_i^{(j)}$ be the maximum likelihood estimator of $\mu_i = Pr\left(Y_i^{(j)} = 1 | X_i^{(j)}\right)$ under the null hypothesis $H_0 : \gamma^{(j)} = 0$. Further, let $X^{(j)} = \left(X_1^{(j)T}, \ldots, X_{n_j}^{(j)T}\right)$ be the $n_j \times k$ matrix of covariates, $G^{(j)} = \left(G_1^{(j)}, \ldots, G_{n_j}^{(j)}\right)^T$ be the genotype vector, $W^{(j)}$ be a diagonal matrix with i^{th} diagonal element $\hat{\mu}_i^{(j)}\left(1 - \hat{\mu}_i^{(j)}\right)$, and $\tilde{G}^{(j)} = G^{(j)} - X^{(j)}\left(X^{(j)T}W^{(j)}X^{(j)}\right)^{-1}X^{(j)T}W^{(j)}G^{(j)}$ be the covariate-adjusted genotype vector. Then, the score statistic for testing $H_0 : \gamma^{(j)} = 0$ will be $S^{(j)} = \sum_{i=1}^{n_j}\tilde{G}_i^{(j)}\left(Y_i^{(j)} - \hat{\mu}_i^{(j)}\right)$. To apply the saddlepoint approximation (SPA)-based score test, we first need to calculate the cumulant generating function (CGF) of the score statistic and its first and

second derivatives given by,

$$K^{(j)}(t) = \sum_{i=1}^{n_j} \log\left(1 - \hat{\mu}_i^{(j)} + \hat{\mu}_i^{(j)} e^{\tilde{G}_i^{(j)} t}\right) - t \sum_{i=1}^{n_j} \tilde{G}_i^{(j)} \hat{\mu}_i,$$

$$K'^{(j)}(t) = \sum_{i=1}^{n_j} \frac{\hat{\mu}_i^{(j)} \tilde{G}_i^{(j)}}{\left(1 - \hat{\mu}_i^{(j)}\right) e^{-\tilde{G}_i^{(j)} t} + \hat{\mu}_i^{(j)}} - \sum_{i=1}^{n_j} \tilde{G}_i^{(j)} \hat{\mu}_i^{(j)}, \quad \text{and}$$

$$K''^{(j)}(t) = \sum_{i=1}^{n_j} \frac{\left(1 - \hat{\mu}_i^{(j)}\right) \hat{\mu}_i^{(j)} \tilde{G}_i^{(j)^2} e^{-\tilde{G}_i^{(j)} t}}{\left[\left(1 - \hat{\mu}_i^{(j)}\right) e^{-\tilde{G}_i^{(j)} t} + \hat{\mu}_i^{(j)}\right]^2}.$$

Using the saddlepoint approximation method (*Barndorff-Nielsen*, 1990; *Daniels*, 1954), the distribution function of $S^{(j)}$ at the observed score statistic can be approximated by,

$$\Pr\left(S^{(j)} < s\right) \approx \Phi\left\{w + \frac{1}{w} \log\left(\frac{v}{w}\right)\right\},$$

where $w = \text{sgn}(\hat{t})\sqrt{2\left(\hat{t}s - K^{(j)}\left(\hat{t}\right)\right)}$, $v = \hat{t}\sqrt{K''^{(j)}\left(\hat{t}\right)}$, $\hat{t}$ is the solution to the equation $K'^{(j)}\left(\hat{t}\right) = s$, and Φ is the standard normal distribution function. The fastSPA (*Dey et al.*, 2017) test implements a faster version of this saddlepoint approximation method, which can be applied to obtain the p value $p^{(j)}$. One of the steps implemented in the fastSPA test is to apply the saddlepoint approximation method only if the score statistic lies outside a certain standard deviation threshold from the mean. If the score statistic lies inside the standard deviation threshold, then the fastSPA test uses the normal approximation to calculate the p values because the normal approximation behaves well near the mean. In this chapter, we will consider the p values using two such standard deviation threshold, 2 and 0.1, and will denote the tests by fastSPA – 2 and fastSPA – 0.1, respectively.

5.2.2 P Value-Based Meta-Analysis and Normal Distribution-Based Z-Score Method

We first introduce a framework for p value-based meta-analysis. In this framework, the study-specific signed p values $\left(p^{(j)}s\right)$ are inverted to obtain the signed scores $R^{(j)}$s using some distributions $F^{(j)}$s, for $j = 1, \ldots, J$, where the signs are determined by the directions of associations. We call $F^{(j)}$s reference distributions. Then, the meta-analysis score is given by $R_{meta} = \sum_{j=1}^{J} R^{(j)}$ where each $R^{(j)} \sim F^{(j)}$ under the null hypothesis of no association. Traditional Z-score-based meta-analysis is a special case of this framework, where the reference distributions are normal distributions with means zero and variances given by the effective sample sizes of the individual studies. The effective sample size (*Han and Eskin*, 2011) is calculated as $n_j^* = 4n_{j1}n_{j0}/n_j$, where n_{j1} and n_{j0} are the number of cases and controls in the j^{th} study, respectively. This meta-analysis method first inverts the p values using a standard normal distribution to obtain the signed Z-scores $Z^{(j)} = \pm\Phi^{-1}\left(p^{(j)}/2\right)$, where the signs depend on the directions of associations. Then, the scores $R^{(j)}$s are calculated as $R^{(j)} = \sqrt{n_j^*}Z^{(j)}$, for $j = 1, \ldots, J$, and the meta-analysis score is given by $R_{meta} = \sum_{j=1}^{J} R^{(j)} \sim N\left(0, \sum_{j=1}^{J} n_j^*\right)$ under the null hypothesis. We can test the null hypothesis of no association between the phenotype and the variant by testing $Z_{meta} = R_{meta}/\sqrt{\sum_{j=1}^{J} n_j^*}$, which follows $N(0,1)$ under the null hypothesis.

This meta-analysis strategy can control for type I error rates when each study-specific p value follows the uniform distribution. When the case-control is unbalanced and variants are rare, however, each study-specific test statistic $S^{(j)}$ can have a discrete and often very skewed null distribution, which can result in the set of possible study-specific p values to be discrete, and the two-sided probabilities that constitute those p values, to be asymmetric. In such situations, although SPA can be applied to control type I error rates within each individual study, inverting such SPA-based p values to normally distributed Z-scores might not be appropriate, and can introduce

systematic biases.

We notice that the best possible reference distribution would be the null distribution of the score statistic $S^{(j)}$ under model (5.1) (let it be $\tilde{F}^{(j)}$). In that case, $R^{(j)}$s will be the same as $S^{(j)}$s. Within each individual study, $\tilde{F}^{(j)}$ can be approximated based on the CGF of the score statistic, using the SPA method. However, it is difficult to share the CGFs as summary level statistics. In our first method, we propose a simpler technique to approximate $\tilde{F}^{(j)}$s using summary level statistics and suggest sharing $S^{(j)}$s instead of the p values so that we can directly use $R^{(j)} = S^{(j)}$. This is equivalent to a p value-based meta-analysis using the approximations of $\tilde{F}^{(j)}$s as the reference distributions $F^{(j)}$s, because $R^{(j)}$s will closely approximate $S^{(j)}$s when $F^{(j)}$s closely approximate $\tilde{F}^{(j)}$. For the second approach, we suggest sharing the overall genotype counts from the individual studies to construct our reference distributions. Although our approaches require more information than just the p values, case-control sample sizes and directions of associations, the additional information is also summary level information and hence does not need individual level data.

5.2.3 CGF Sharing-Based Method

If studies share the observed score statistic $S^{(j)}$s and their corresponding CGF $K^{(j)}$s under the null distribution, then the meta-analysis score and its CGF can be calculated as $R_{meta} = \sum_{i=1}^{J} S^{(j)}$ and $K_{meta} = \sum_{i=1}^{J} K^{(j)}$, respectively. The saddlepoint approximation can be applied on R_{meta} to obtain the meta-analysis p value. Since it is difficult to share complicated functions like $K^{(j)}$s using summary statistics, studies can only share the functions at some pre-specified node values and reconstruct the functions at the meta-analysis stage using spline approximations.

Notice that the CGFs and their derivatives $K^{(j)}, K'^{(j)}, K''^{(j)}$s are smooth functions as evident from their algebraic expressions. Therefore, cubic splines (*Bartels et al.*, 1987; *Press et al.*, 1992) should provide good approximations of these functions. We

provided some examples of these functions and their spline approximations under different case-control ratios and allele frequencies in Appendix F (Figures F.1,F.2 and F.3) Further, if we apply cubic splines to approximate one of these functions in the meta-analysis stage, the other two functions can be obtained through algebraic or numerical differentiations and integrations. For example, if we obtain the cubic spline approximation $\hat{K}'^{(j)}$ of $K'^{(j)}$, we can easily calculate the derivative $\hat{K}''^{(j)}(t)$ and the integral $\hat{K}^{(j)}(t)$ at any t through either algebraic or numerical differentiation and integration, as $\hat{K}'^{(j)}$ is a piece-wise cubic smooth polynomial.

For our purpose, we approximate $K'^{(j)}$s using cubic splines because $K'^{(j)}$, being involved in the saddlepoint equation, is the most important function of these three. To obtain more accurate approximation of both $K'^{(j)}$ and $K''^{(j)}$, we use cubic Hermite splines (*Bartels et al.*, 1987; *Kreyszig*, 2006) instead of cubic natural splines. Hermite spline method takes the values of a function (in our case $K'^{(j)}$) and its derivative (in our case $K''^{(j)}$) at some node points, and provides a piece-wise cubic smooth approximation $\left(\hat{K}'^{(j)}\right)$ where both the functional values as well as the derivative values are preserved at the node points. This means, not only does $\hat{K}'^{(j)}$ match with $K'^{(j)}$ at the node points, but also the derivative of $\hat{K}'^{(j)}$ matches with $K''^{(j)}$ at those node points. Therefore, if an individual study shares the functional values of $K'^{(j)}$s and $K''^{(j)}$s at some pre-specified node points, Hermite spline method can fit both of these functions simultaneously. On the other hand, cubic natural spline can only preserve the functional values of the function it approximates. This phenomenon is illustrated in Figures F.1,F.2 and F.3 using some examples with different case-control ratios and MAFs. For these illustrations, the K' functions were fitted using spline approximations on seven node points, and $\hat{K}''$s and $\hat{K}$s were calculated using algebraic differentiations and numerical integrations, respectively. Nodes were selected based on the algorithm we discussed later. The examples show that the cubic natural splines only fit the K' functions at the node points, but the algebraic differentiations of $\hat{K}'$

functions can result in poor approximations of the K'' functions. On the other hand, the cubic Hermite splines fit both the K' and K'' functions simultaneously at the node points.

To calculate the Hermite spline approximations, suppose the values of $K'^{(j)}$ and $K''^{(j)}$ are provided at the node points $t_0 < t_1 < \ldots < t_r$. Then the Hermite spline interpolation of $K'^{(j)}$ between two node points $[t_k, t_{k+1}]$ is given by,

$$\hat{K}'^{(j)}(t) = h_{00}(x)K'^{(j)}(t_k) + h_{10}(x)(t_{k+1} - t_k)K''^{(j)}(t_k) + h_{01}(x)K'^{(j)}(t_{k+1}) + h_{11}(x)(t_{k+1} - t_k)K''^{(j)}(t_{k+1}),$$

where $x = (t - t_k)/(t_{k+1} - t_k)$, and h_{00}, h_{01}, h_{10}, and h_{11} are the Hermite basis functions, $h_{00} = (1+2t)(1-t)^2$, $h_{10} = t(1-t)^2$, $h_{01} = t^2(3-2t)$, $h_{11} = t^2(t-1)$. Linear extrapolation is applied to obtain the functional values outside the boundaries t_0 and t_r, using the slopes at those boundaries. The algorithm that we implemented in our R package to obtain the optimal set of nodes, is discussed in Appendix F.

Once we obtain $\hat{K}'^{(j)}$, we can algebraically differentiate the function to get $\hat{K}''^{(j)}$,

$$\hat{K}''^{(j)}(t) = \frac{dh_{00}(x)}{dt}K'^{(j)}(t_k) + \frac{dh_{10}(x)}{dt}(t_{k+1} - t_k)K''^{(j)}(t_k) + \frac{dh_{01}(x)}{dt}K'^{(j)}(t_{k+1}) + \frac{dh_{11}(x)}{dt}(t_{k+1} - t_k)K''^{(j)}(t_{k+1}).$$

Similarly, an algebraic or numerical integration can be performed to obtain $\hat{K}^{(j)}$. The constant of integral can be determined using the initial condition $\hat{K}^{(j)}(0) = 0$. Then, the CGF of the meta-analysis score and its derivatives can be approximated by $\hat{K}_{meta} = \sum_{j=1}^{J} \hat{K}^{(j)}$, $\hat{K}'_{meta} = \sum_{j=1}^{J} \hat{K}'^{(j)}$, and $\hat{K}''_{meta} = \sum_{j=1}^{J} \hat{K}''^{(j)}$.

5.2.4 Genotype Count-Based Method

The aforementioned CGF sharing method requires to share Hermite spline nodes to construct the CGFs in the meta-analysis stage. But in many situations, this information would not be available, especially when the meta-analysis is conducted

using publicly available summary statistics. Here we propose a practical alternative approach to approximate the CGFs using the genotype counts (number of 0, 1, 2 genotypes) at different markers. Genotype counts are more readily available and software independent than the information required to be shared in the spline-based method. For rare variants, where homozygous minor genotypes are usually not present in the data, or for variants that follow Hardy-Weinberg equilibrium, sharing the minor allele counts (MAC) will be sufficient, as the genotype counts can be easily calculated based on the MACs.

Suppose, for the j^{th} study, the genotype counts for the variant to be tested are m_{j0}, m_{j1} and m_{j2} corresponding to the genotypes 0,1 and 2 respectively. Then, we can construct the genotype vector $G^{(j)^*}$ of length n_j where the first m_{j2} elements are 2s, next m_{j1} elements are 1s, and the rest are 0s. We propose using the null distribution (let it be $F^{(j)^*}$) of the score statistic in the following genotype-only model (5.2) as our reference distribution,

$$\text{logit}\left[\Pr\left(\mathrm{Y_i^{(j)}} = 1 | \mathrm{G_i^{(j)^*}}\right)\right] = \alpha^{(\mathrm{j})^*} + \mathrm{G_i^{(j)^*}}\gamma^{(\mathrm{j})^*} \tag{5.2}$$

where $G_i^{(j)^*}$ is the i^{th} elements of $G^{(j)^*}$, $\alpha^{(j)^*}$ is the intercept and $\gamma^{(j)^*}$ is the genotype log odds ratio. Intuitively, when the non-genetic covariates are relatively balanced across cases and controls, the discreteness and asymmetry in the null distribution of the score statistic mainly depend on the imbalance or the rarity of the phenotype and the genotype. Therefore, the null distribution of the score statistic under the genotype-only model can be a reasonable alternative to the traditionally used normal distribution, as a reference distribution. To apply this method, we first need to calculate the CGF of the score statistic and its first and second derivatives in the

genotype-only model (5.2) given by,

$$K^{(j)^*}(t) = \sum_{i=1}^{n_j} \log\left(1 - \hat{\mu}^{(j)^*} + \hat{\mu}^{(j)^*} e^{\bar{G}_i^{(j)^*} t}\right),$$
$$K'^{(j)^*}(t) = \sum_{i=1}^{n_j} \frac{\hat{\mu}^{(j)^*} \bar{G}_i^{(j)^*}}{\left(1 - \hat{\mu}^{(j)^*}\right) e^{-\bar{G}_i^{(j)^*} t} + \hat{\mu}^{(j)^*}}, \quad \text{and}$$
$$K''^{(j)^*}(t) = \sum_{i=1}^{n_j} \frac{\left(1 - \hat{\mu}^{(j)^*}\right) \hat{\mu}^{(j)^*} \bar{G}_i^{(j)^* 2} e^{-\bar{G}_i^{(j)^*} t}}{\left[\left(1 - \hat{\mu}^{(j)^*}\right) e^{-\bar{G}_i^{(j)^*} t} + \hat{\mu}^{(j)^*}\right]^2},$$

where $\bar{G}_i^{(j)^*} = G_i^{(j)^*} - \bar{G}^{(j)^*}$ is the mean-centered genotypes, and $\hat{\mu}^{(j)^*}$ = the proportion of cases, is the maximum likelihood estimator of $\mu^{(j)^*} = \Pr\left(Y_i^{(j)} = 1\right)$ under the null hypothesis $H_0^* : \gamma^{(j)^*} = 0$. Based on this CGF, we can approximate the distribution $F^{(j)^*}$ and calculate the score $R^{(j)}$ by inverting $F^{(j)^*}$ at the signed fastSPA p value $\pm p^{(j)}$, which is calculated from the model (5.1) with all covariates. Since the signed p values have one-to-one relationships with the score values, the inversion of $\pm p^{(j)}$ to obtain the score $R^{(j)}$ can be performed using root-finding algorithms such as Newton-Raphson (*Press et al.*, 1992), Brent (*Brent*, 1973), bisection (*Press et al.*, 1992) etc. In our implementation, we applied Brent's method for this purpose. The meta-analysis score $R_{meta} = \sum_{j=1}^{J} R^{(j)}$ will then have the CGF $K_{meta} = \sum_{j=1}^{J} K^{(j)^*}$, and we can apply the SPA test on R_{meta} to obtain the meta-analysis p value.

5.3 Numerical Simulations

We evaluated the type I error rates and empirical powers of the Z-score-based and proposed methods through extensive simulation studies. We considered three different simulation study settings. For the first setting, we meta-analyzed seven studies coming from the same population where the genotypes and the non-genetic covariates are simulated independently. For the second setting, we considered a meta-analysis of seven studies where the genotypes and the non-genetic covariates were

simulated based on the MAF and principal component (PC) scores in different ethnic groups in the UK Biobank data. In the third setting, we assessed the performance of the methods when a smaller but balanced case-control study is meta-analyzed along with a small number of larger but unbalanced biobank-based studies.

5.3.1 Simulation Study 1 : Meta-Analyzing Seven Studies from the Same Population

Our first simulation study was designed to represent a meta-analysis of multiple studies from the same population. We considered seven studies with sample sizes $n_j = 2000$ for all $j = 1, \ldots, 7$. We further considered three case-control ratios: balanced with the case-control ratio of $1:1$ within each study, moderately unbalanced with the case-control ratio of $1:9$ within each study, and extremely unbalanced with the case-control ratio of $1:49$ within each study. For each choice of case-control ratio, the phenotypes in the j^{th} study were simulated using the following logistic model,

$$\text{logit}\left[\Pr\left(Y_i^{(j)} = 1\right)\right] = \alpha^{(j)} + 0.5 \times \left(X_1^{(j)} + X_2^{(j)}\right) + G_i^{(j)}\gamma^{(j)}, \tag{5.3}$$

for $i = 1, \ldots, n_j$, where $X_1^{(j)} \sim N(0,1)$ and $X_2^{(j)} \sim$ Bernoulli(0.5) were the non-genetic covariates, and the genotypes ($G_i^{(j)}$s) were generated from a Binomial$(2, \text{p})$ distribution where p (same across the seven studies) was the minor allele frequency (MAF). The intercepts ($\alpha^{(j)}$s) were selected such that the prevalence within each study would become 0.01. The parameters $\gamma^{(j)}$s represent the within-study log-odds ratios. For the type I error comparisons, all $\gamma^{(j)}$s were set to be 0. A wide range of $\gamma^{(j)}$ values were used for the power calculations (see Section 5.4).

To compare the type I error rates of different methods under different MAFs, we considered five different MAFs, $p = 0.001, 0.005, 0.01, 0.05, 0.1$, and simulated 5×10^8 variants for each of the MAFs and the three case-control ratios. We recorded the

number of rejections at $\alpha = 5 \times 10^{-5}$ and 5×10^{-8} genome-wide significance levels. We further performed a power comparison with 5000 simulated variants for each of the three case-control ratios and two choices of the MAF $p = 0.01$, and 0.05, at different values of $\gamma^{(j)}$. As the genome-wide significance threshold for power calculations, we used both a nominal $\alpha = 5 \times 10^{-8}$, and a type I error adjusted empirical *alpha* where the corresponding method has type I error 5×10^{-8}. The empirical α level was calculated based on 5×10^8 simulated datasets from the simulation setting described above (seven studies, each with 2000 samples) where the MAFs were sampled from the MAF spectrum (Figure I.1) of the White British ancestry group ($\sim$ 117K samples) in the UK Biobank interim release data.

5.3.2 Simulation Study 2 : Trans-Ethnic Meta-Analysis of Seven Studies

Our second simulation study was designed to represent a trans-ethnic meta-analysis where the MAFs can be different across the studies. We considered seven studies with sample sizes $n_j = 2000$ for all $j = 1, \ldots, 6$, and $n_7 = 1500$. To simulate the genotypes and the non-genetic covariates from a realistic meta-analysis of GWAS, we used genotype data from the UK Biobank interim release data (*UK Biobank*, 2015). The first five studies included first four principal component (PC) scores as covariates and genotypes simulated from the MAF spectrum of the White ancestry group in the UK Biobank samples. To maintain the correlated nature of the genotypes and the PC scores, genotypes were simulated using PC scores. We further added a binary covariate generated from a Bernoulli(0.5) distribution independent of the PC scores and the genotypes. Covariates and genotypes were simulated in a similar way for study six and seven based on the South Asian and Black ancestry groups, respectively. The model to simulate the phenotypes was similar to the one used in the first simulation study, except for different non-genetic covariates. Detailed explanation of the simulation procedure is provided in Appendix G.

In trans-ethnic studies, variants have different MAFs across different ancestry groups. To calculate the type I error rates for diverse scenarios of MAFs, we first considered three MAF bins for the alleles of the simulated variants: rare variants with MAF ≤ 0.01, low frequency variants with $0.01 < \text{MAF} \leq 0.05$ and common variants with MAF > 0.05. We then categorized the simulated variants in four categories based on their allele frequencies (AF): a) all rare, when the variant has the same minor rare allele in all seven studies, b) all low frequency, when the variant has the same low frequency allele in all seven studies, c) all common, when the variant has the same common allele in all seven studies, and d) different AF, when the variant falls in different MAF bins in at least two different studies. The different AF category also includes variants which have different alleles as the minor alleles in different studies. For each variant category and case-control ratio, we simulated 5×10^8 datasets under the null hypothesis and recorded the number of rejections at the genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}.

5.3.3 Simulation Study 3 : Meta-Analyzing a Balanced Case-Control Study with Two Larger Unbalanced Studies

We investigated the performance of different meta-analysis strategies when a balanced case-control study, which is smaller in sample size, is meta-analyzed along with two larger biobank-based unbalanced studies. This simulation study represents the real world meta-analyses where the researchers collect balanced case-control data on rare traits/diseases, and attempt to meta-analyse them with association results from a small number of larger cohort-based studies. To simulate the genotypes, non-genetic covariates and the phenotypes, we used the same simulation and logistic regression models as in our first simulation study setting. The sample size for the balanced case-control study was 2000 with 1000 cases and 1000 controls, and the unbalanced studies had sample size 10000 each. We considered two case-control ratios for these

unbalanced studies: moderately unbalanced with case : control = 1 : 9 within each study, and extremely unbalanced with case : control = 1 : 49 within each study. For each of the case-control ratio, we compared the type I error rates of different methods under five different MAFs, $p = 0.001, 0.005, 0.01, 0.05, 0.1$ based on 5×10^8 simulated variants each.

For the first two simulation settings and the unbalanced studies in the third simulation setting, the within-study p values were calculated using the traditional score test (Score), fastSPA test with 2 standard deviations threshold (fastSPA – 2), and fastSPA test with 0.1 standard deviations threshold (fastSPA – 0.1). Since score test is relatively well-calibrated for balanced case-control studies, only Score p values were calculated for the balanced study in the third simulation setting. We then considered the following meta-analysis methods to compare their type I error rates and empirical powers: Z-score-based meta-analysis (Z-score), CGF sharing-based meta-analysis (CGF-Spline), and genotype count sharing-based meta-analysis (GC). Score p values were meta-analyzed using the Z-score method, fastSPA – 2 and fastSPA – 0.1 p values were meta-analyzed using the Z-score and GC methods, and the within-study observed score statistics were meta-analyzed using the CGF-Spline method. For the balanced case-control study in the third simulation setting, the Z-scores obtained from the Score p values were used in the GC method, and the corresponding normal distribution-based CGFs were used in the CGF-Spline method. We also compared the type I error rates and the empirical powers of a joint analysis (Joint) using the fastSPA – 2 test on the pooled data as the gold standard.

To assess the scalability of our proposed methods in a realistic GWAS meta-analysis scenario, we calculated their computation times based on 10^4 simulated variants under the null hypothesis in the second simulation setting. We also calculated the computation time required to prepare the summary information for the CGF sharing-based meta-analysis method using a single study of 2000 samples and 10^4

simulated variants where the PC scores and MAFs were sampled based on the White ancestry group of the UK Biobank data.

5.4 Results

In this section, we evaluate the performance of the proposed methods against the Z-score-based meta-analysis based on the numerical simulations described above.

5.4.1 Type I Error Comparison

The type I error comparison based on simulation study 1 (Figure 5.2) clearly shows that the proposed CGF-Spline and GC methods provided well-controlled type I error rates across all the MAFs and all the case-control ratios. Expectedly, the joint analysis also controlled the type I error rates. On the other hand, the Z-score method resulted in inflated type I error rates in moderately unbalanced and extremely unbalanced settings, especially for the rarer minor allele frequencies. Interestingly, the Z-score method with fastSPA-0.1 performed worse than that with fastSPA-2, although fastSPA-0.1 used the saddlepoint approximation to more variants. This further verifies our assertion that using normal distributions to invert the study-specific p values which are possibly discrete, asymmetric and originally calculated using the saddlepoint approximation, can result in failure to control type I error in the meta-analysis process. For MAF $= 0.001$ under the extremely unbalanced setting, there is conservative behavior shown by the Z-score method when using fastSPA – 0.1 or fastSPA – 2 p values at $\alpha = 5 \times 10^{-5}$ level (empirical type I error rates 3.55×10^{-6} and 2.76×10^{-5}, respectively). All methods provided well-controlled type I error rates for the balanced case-control ratio.

Similar observation follows for simulation study 2. The type I error comparison (Figure 5.3) suggests that our proposed methods showed no sign of type I error inflation across different MAFs and case-control ratios, whereas the Z-score method

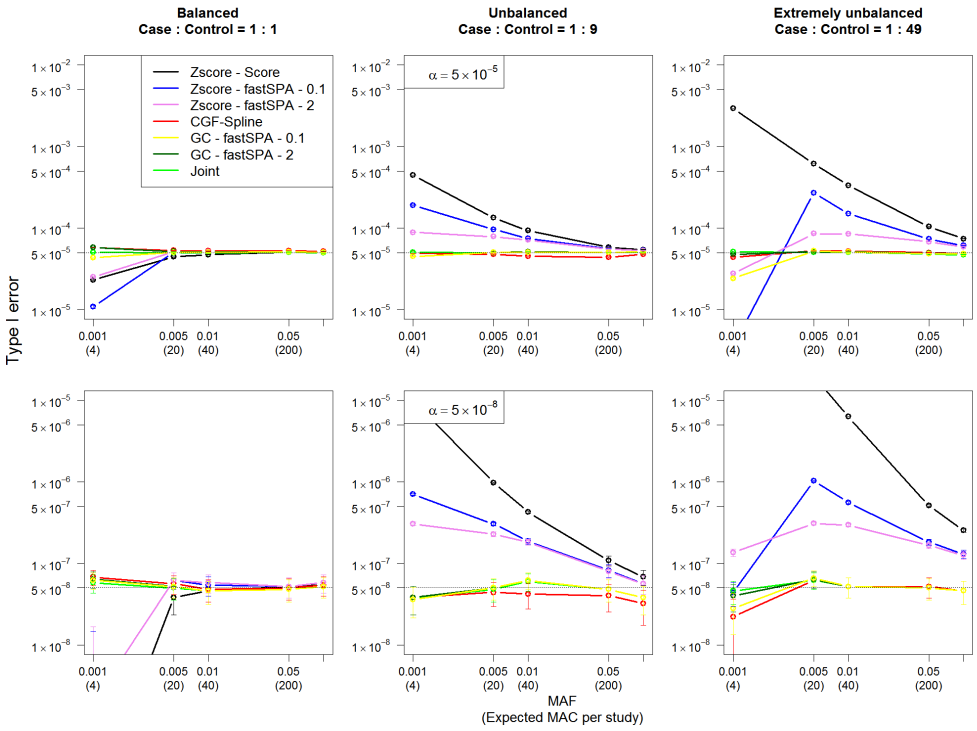


Figure 5.2: Type I error comparison among different meta-analysis methods and joint analysis, in simulation study 1. Joint represents the joint analysis with the pooled data. The top and the bottom panels show empirical type I error rates at genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. From left to right, the plots consider the within-study case-control ratios $1:1$, $1:9$ and $1:49$, respectively. In each plot, the X-axis represents MAFs with expected MACs in parenthesis, and the Y-axis (in logarithmic scale) represents the empirical type I error rates. 95% confidence intervals at different MAFs are also presented.

resulted in inflated type I error rates for the moderately unbalanced and extremely unbalanced settings, especially for the all rare, all low frequency and different MAF categories. Z-score method using Score p values had the maximum inflation across all categories.

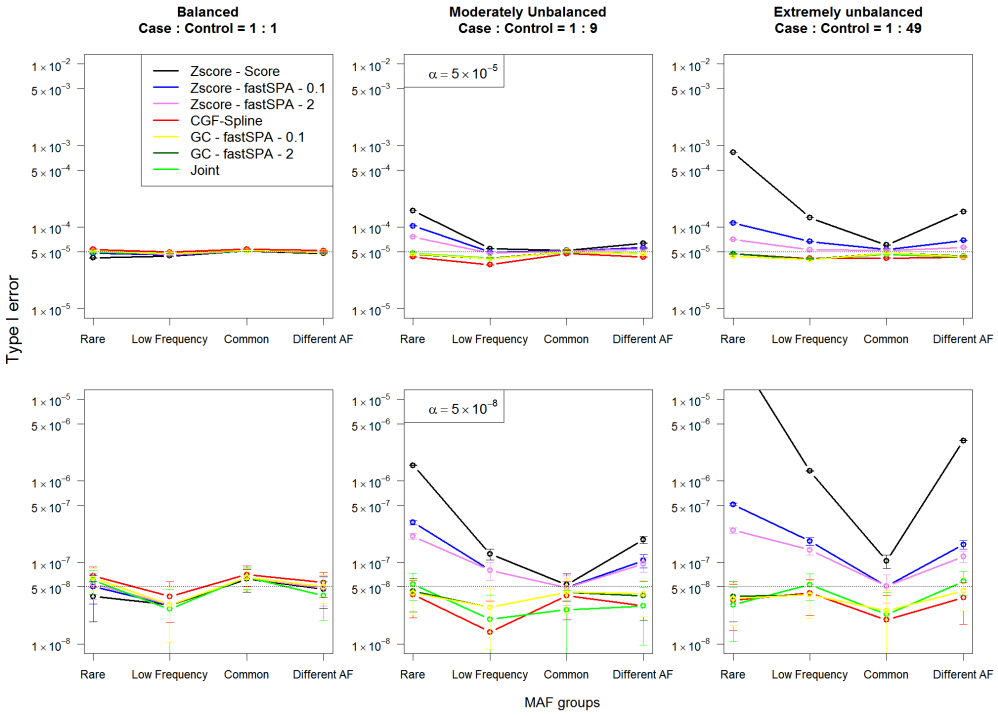


Figure 5.3: Type I error comparison among different meta-analysis methods and joint analysis, in simulation study 2. Joint represents the joint analysis with the pooled data. The top and the bottom panels show empirical type I error rates at genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. From left to right, the plots consider the within-study case-control ratios $1:1$, $1:9$ and $1:49$, respectively. In each plot, the X-axis represents different MAF groups: Rare (variant is rare in all studies), Low frequency (variant is low frequency in all studies), Common (variant is common in all studies) and Different AF (variant is in different allele frequency group in at least two different studies). The Y-axis (in logarithmic scale) represents the empirical type I error rates. 95% confidence intervals at different MAFs are also presented.

In simulation study 3, we also have similar results (Figure 5.4) for our proposed methods. However, the Z-score method using the fastSPA – 0.1 or fastSPA – 2 p values showed no sign of significant type I error inflation in the extremely unbalanced

case-control setting, and only slight inflation in the moderately unbalanced setting. The maximum empirical type I error for the Z-score-based meta-analysis with fastSPA p values was 8.01×10^{-8} (1.6 times the nominal $\alpha = 5 \times 10^{-8}$), observed at MAF = 0.005 in the moderately unbalanced setting. This suggests that the Z-score-based method can be adequate for controlling the type I error rates when only a small number of biobank-based studies are included in the meta-analysis. However, as seen from the other two simulation studies, the Z-score method may fail to control type I error rates when large number of unbalanced studies are involved.

5.4.2 Power Comparison

Next, we compare the empirical powers of different meta-analysis strategies alongwith the joint analysis as the gold standard under the first simulation setting. Because the Z-score-based meta-analysis method provided inflated type I error rates as seen in the type I error comparisons, we used empirical α levels calculated from type I error simulations for each method where the empirical type I error rate becomes 5×10^{-8}. The power curves (Figure 5.5) show that the Z-score method has slightly lower power (lowest when using score test p values) in the moderately and extremely unbalanced case-control ratios. Our proposed methods provide very similar power to the joint analysis, and all methods provide similar power in the balanced case-control setting. When nominal $\alpha = 5 \times 10^{-8}$ level was used (Figure I.2), the Z-score method expectedly showed higher powers in the unbalanced settings since it is not calibrated for its type I errors.

5.4.3 Computation Times of the Proposed Methods

Figure 5.6 shows the projected computation times of our proposed methods for meta-analyzing 10 million variants across seven studies at different case-control ratios. The results suggest that both methods are scalable for GWASs, with the longest

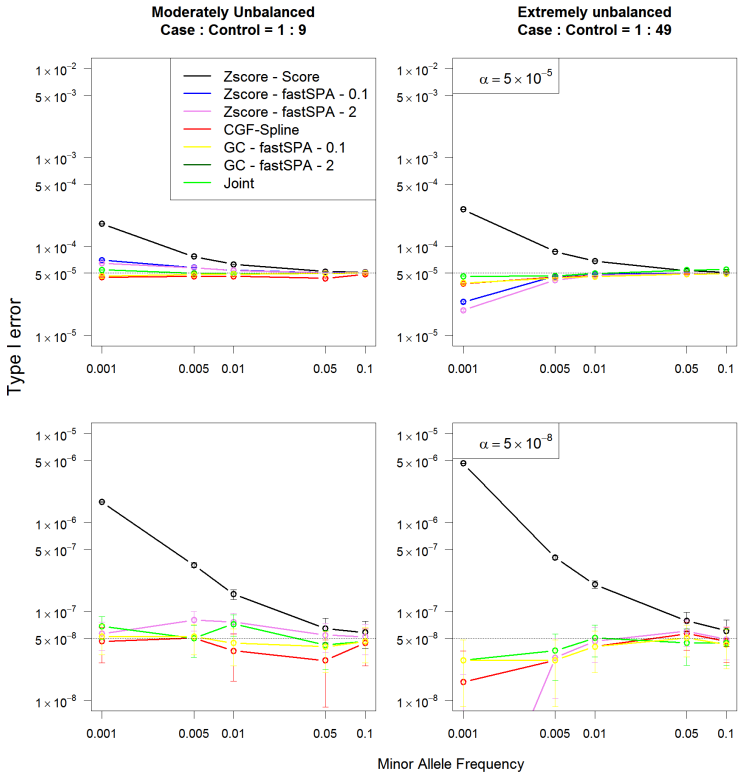


Figure 5.4: Type I error comparison among different meta-analysis methods and joint analysis, in simulation study 3. Joint represents the joint analysis with the pooled data. The top and the bottom panels show empirical type I error rates at genome-wide significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8}, respectively. The left and right panels consider the within-study case-control ratios $1 : 9$ and $1 : 49$, respectively for the unbalanced studies. In each plot, the X-axis represents MAFs with expected MACs in parenthesis, and the Y-axis (in logarithmic scale) represents the empirical type I error rates. 95% confidence intervals at different MAFs are also presented. The empirical type I error rates were almost identical between ZScore – fastSPA – 2 and ZScore – fastSPA – 0.1, and between GC – fastSPA – 2 and GC – fastSPA – 0.1, and hence the lines are sometimes overlapped in this plot.

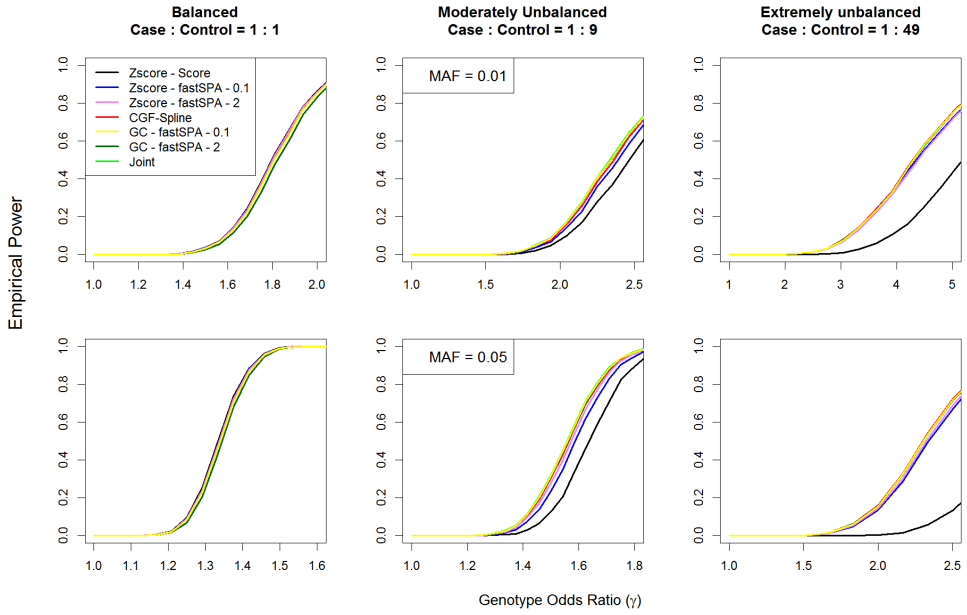


Figure 5.5: Power curves for the Z-score, CGF-Spline and Genotype Count (GC) methods. Top panel considers MAF = 0.01 and bottom panel considers MAF = 0.05. From left to right, the plots consider case-control ratios 1 : 1, 1 : 9 and 1 : 49, respectively. In each plot the X-axis represents genotype odds ratios and the Y-axis represents the empirical power. Empirical power was estimated from 5000 simulated datasets at their type I error adjusted empirical α levels where their empirical type I errors are equal to 5×10^{-8}.

observed computation time of 59.4 CPU-hours for the GC method using fastSPA – 0.1 p values in the balanced case-control setting. The GC method, when using the fastSPA – 2 p values, shows similar computation times as the CGF-Spline method (3 — 5 CPU-hours). We note that the GC method does not require any additional computation within the individual studies whereas the CGF-Spline method involves a node-finding step within each study. We calculated the additional computation time required within each study to prepare the summary information for the CGF-Spline method. The projected additional computation times for preparing the summary information of 10 million variants in a single study of 2000 samples were 4.1 CPU-hours for case-control ratio 1 : 1, 4.3 CPU-hours for case-control ratio 1 : 9, and 4.4 CPU-hours for case-control ratio 1:49.

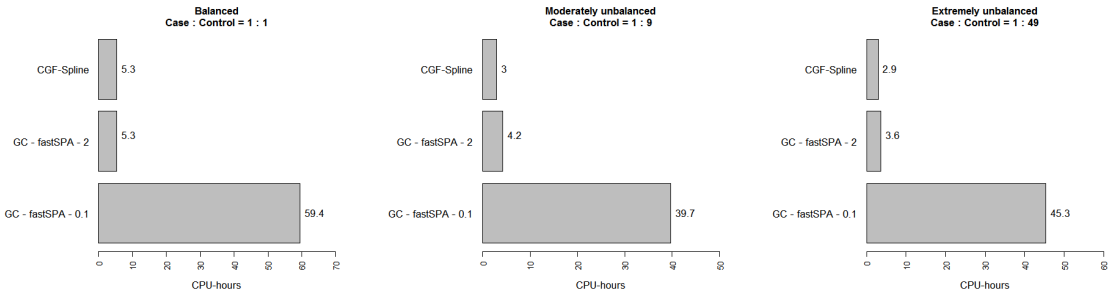


Figure 5.6: Projected computation times of our proposed methods for meta-analyzing seven studies at 10 million variants as described in simulation study 2. The computation times are based on a meta-analysis of 10000 simulated variants on an Intel i7 2.70GHz processor, and then projecting it onto 10 million variants.

5.5 UK Biobank Data Analysis

We demonstrated the performance of our proposed methods by analyzing two phenotypes based on the UK Biobank interim release data (*UK Biobank*, 2015). The UK Biobank (*Bycroft et al.*, 2017) contains detailed phenotypic information based on electronic health records for $\sim$ 500K individuals in the United Kingdom. In the

interim release (May 2015), information on $\sim$ 150K individuals were released to the public. Details about the data and pre-processing are provided in Appendix H. A histogram of the case-control ratios (Figure 5.1) of different binary phenotypes shows that the ratios are heavily skewed towards zero, which means the binary phenotypes are mostly unbalanced.

To compare our proposed methods with the Z-score-based meta-analysis method, we analyzed two phenotypes, Ulcerative Colitis (PheWAS code: 555.2, case : control = 1 : 100), and Psoriasis (PheWAS code: 696.4, case : control = 1 : 165) based on 117,513 unrelated samples from the White British ancestry group of the interim release data. The samples were then divided into 22 groups based on the assessment center where they first consented to be included in the biobank. We selected 19 centers (Table I.1) with at least 5 cases for each of the two phenotypes, and treated these centers as our individual studies to perform association analyses of the phenotypes on the autosomal variants within each of them. For the within-study association analyses, we applied Score, fastSPA – 2 and fastSPA – 0.1 tests, adjusting for age, sex, and first four principal components. Individuals which had phenotype or at least one covariate information missing, were removed from the analysis of that corresponding phenotype. We only applied the within-study tests for variants with within-study MAC at least three. Because the genotype count-based meta-analysis requires the overall genotype counts, we applied our within-study tests on the best called genotypes instead of dosages in the imputed data. We then meta-analyzed the results using the Z-score-based meta-analysis (Z-score), CGF sharing-based meta-analysis (CGF-Spline), and genotype count sharing-based meta-analysis (GC). The meta-analysis methods were only applied for variants that were tested in at least two different studies, and the overall MACs were at least ten. For each phenotype, $\sim$ 29 million variants were meta-analyzed.

The quantile-quantile (QQ) plots presented in Figure 5.7 and Figure 5.8 show that

the meta-analysis p values from our proposed methods closely follow the uniform distribution, whereas those from the Z-score method are either much smaller (Z-score method using Score or fastSPA – 0.1 p values) or larger (Z-score method using fastSPA – 2 p values) than expected for rare variants (MAF < 0.01). This suggests conservative behavior of the Z-score method when using the fastSPA – 2 p values, and extremely anti-conservative behavior when using fastSPA – 0.1 or Score p values. On the other hands, both the CGF-Spline and GC methods improve the accuracy of the meta-analysis p values and provide well-calibrated QQ plots. We also presented the genomic control inflation factors (λ) of different meta-analysis strategies in Table I.2. For Ulcerative Colitis, all our proposed methods showed no inflation or deflation in the genomic controls at p value quantiles $q = 0.01$ and 0.001, whereas the Z-score method showed severely inflated inflation factors when using the Score (eg. $\lambda = 1.33$ at $q = 0.01$) and fastSPA – 0.1 (eg. $\lambda = 3.18$ at $q = 0.01$) p values and deflated inflation factors when using the fastSPA – 2 (eg. $\lambda = 0.82$ at $q = 0.01$) p values at those p value quantiles. This result further supports the observations made from the QQ plots. When considering the inflation factors at the median p value quantile ($q = 0.5$), the CGF-Spline ($\lambda = 0.74$) and GC method using fastSPA – 2 p values ($\lambda = 0.68$) showed deflated inflation factors, and GC method using fastSPA – 0.1 p values ($\lambda = 1.40$) showed inflated inflation factor. This is expected, since the fastSPA p values near the median are not calculated using the saddlepoint approximation as discussed in Chapter IV. In that paper, they also found inflated genomic control factors for fastSPA – 0.1 and deflated genomic control factors for fastSPA – 2 p values at the median level for extremely unbalanced case-control ratios. The inflation factors showed similar patterns for Psoriasis. However, at p value quantile $q = 0.001$, the GC method using fastSPA – 2 p values ($\lambda = 1.09$), and the CGF-Spline method showed slightly larger than expected inflation factors ($\lambda = 1.10$). This might be due to the presence of the Major Histocompatibility Complex (MHC) in the 6p21 region which

contains a large number of polymorphic variants and it is a known associated region for Psoriasis (*Stuart et al.*, 2015). After excluding the MHC region from the inflation factor calculation, the inflation factors became very close to unity.

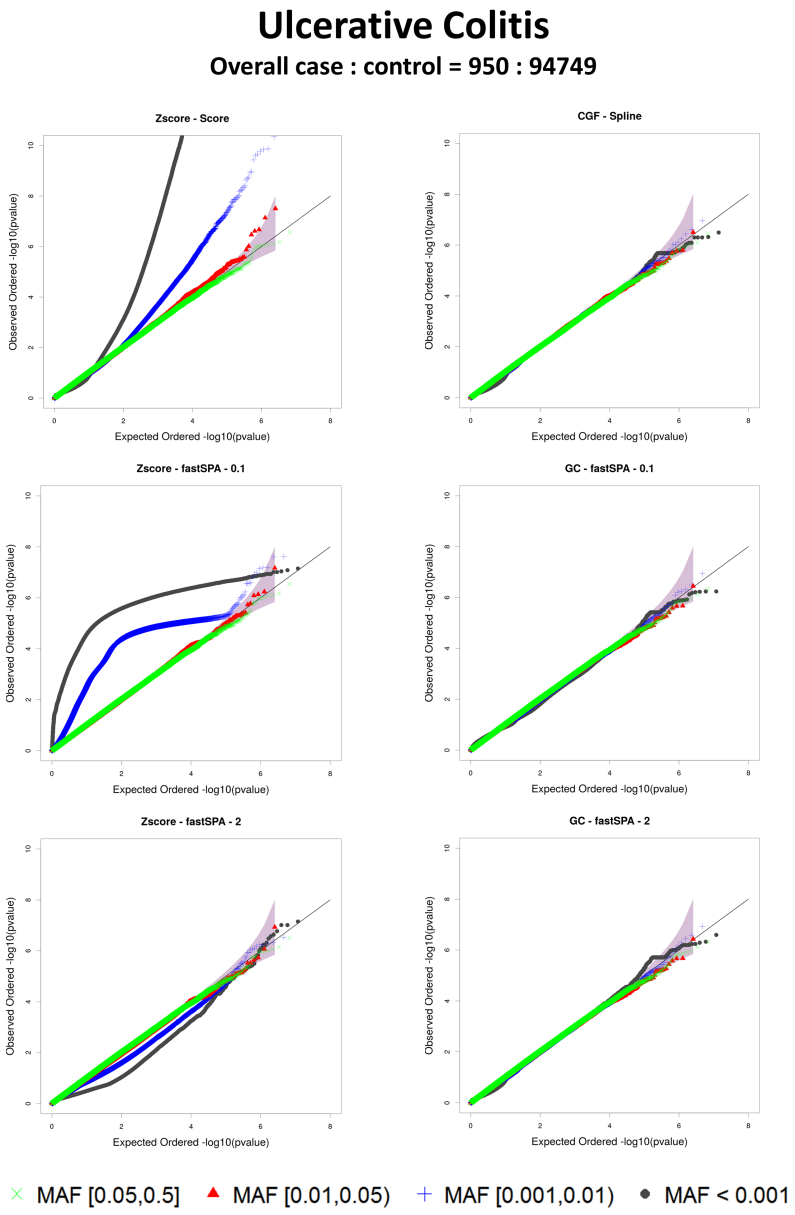


Figure 5.7: Meta-analysis QQ plots for Ulcerative Colitis based on the UK Biobank interim release data. QQ plots using the Z-score method are provided in the left panel, and the QQ plots using our proposed methods are provided on the right panel. Known associated loci in the MHC region were removed from the QQ plots. The plots are color-coded based on different MAF categories.

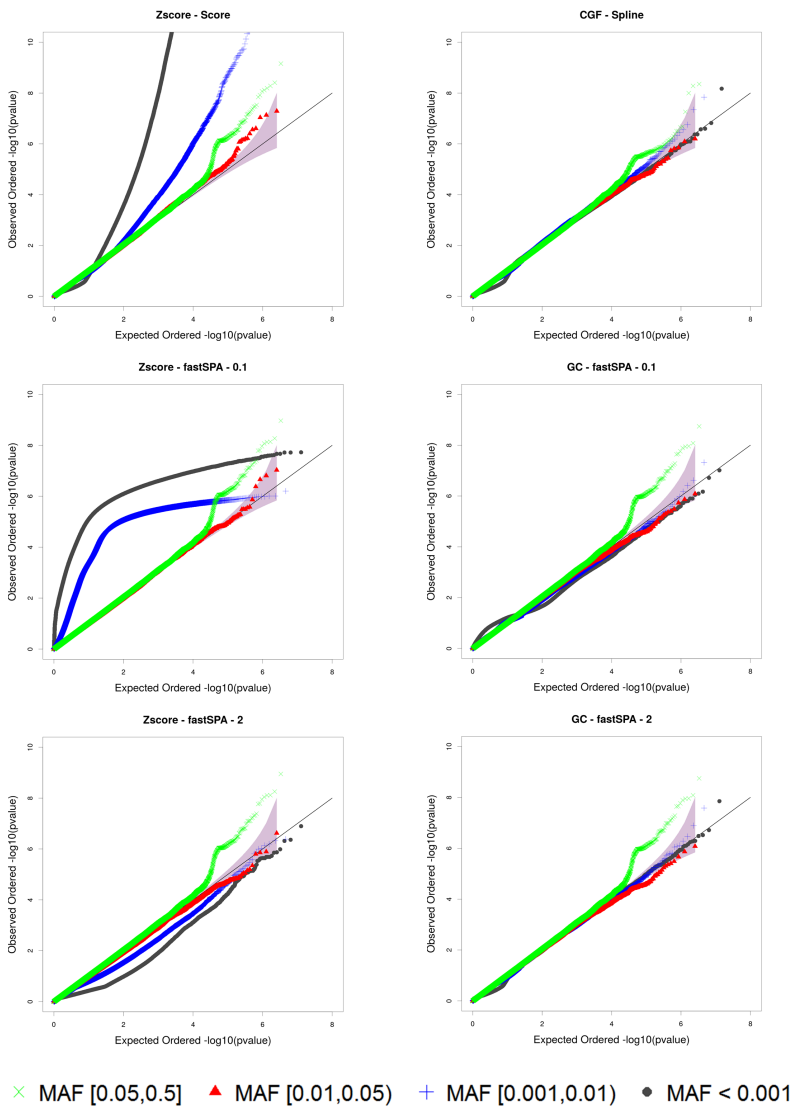


Figure 5.8: Meta analysis QQ plots for Psoriasis based on the UK Biobank interim release data. QQ plots using the Z-score method are provided in the left panel, and the QQ plots using our proposed methods are provided on the right panel. Known associated loci in the MHC region were removed from the QQ plots. The plots are color-coded based on different MAF categories.

The top genome-wide significant SNPs in different regions, identified by the CGF-Spline method, are listed in Table 5.1. For the genotype count method, the top significant SNPs were identical and the corresponding p values were very similar, and

hence those are omitted from the table.

Phenotype	Location	dbSNP ID	Nearest Gene	Alleles	MAF	p value	Previous Findings
Ulcerative Colitis	6:32433654	**rs9268944**	*HLA-DRA*	C>T	0.42	7.49×10^{-12}	*Anderson et al.* (2011)
Psoriasis	6:31251924	**rs12189871**	*HLA-C*	C>T	0.09	9.72×10^{-53}	*Stuart et al.* (2015)
	15:50785016	**rs148783236**	*USP8*	A>G	0.12	1.68×10^{-47}	*Verma et al.* (2018)
	22:21190325	rs549956609	*PI4KA*	G>A	0.0009	6.72×10^{-9}	Potentially novel
	5:158826357	rs918519	*IL12B*	G>A	0.23	5.48×10^{-9}	*Stuart et al.* (2015)
	22:50110828	rs560106765	*BRD1*	A>C	0.0014	1.46×10^{-8}	Potentially novel

Table 5.1: Genome-wide significant ($\alpha = 5 \times 10^{-8}$) SNP-phenotype associations based on the meta-analysis using the CGF-Spline method. The SNPs which are also significant at the corresponding Bonferroni correction level, are shown using their dbSNP ID in bold font. The Bonferroni correction levels were 1.78×10^{-9} for Ulcerative Colitis, and 1.72×10^{-9} for Psoriasis.

To assess the performance of our methods with genotype dosage data, we further performed our within-study tests to calculate the p values, scores and spline-based summary statistics using the dosage data, and then meta-analyzed the results using our proposed methods. For the GC method, we calculated the within-study p values based on the dosages, but constructed the genotype-only model using genotype counts of the best-called genotypes. The resulting QQ plots (Figure I.3) showed no sign of inflation or deflation for our methods, which suggests that the methods are robust for the analysis of dosage data.

5.6 Discussion

We evaluated the performance of the traditional Z-score-based meta-analysis strategy to combine association results from multiple unbalanced genome-wide association studies, and proposed two alternative strategies that can provide well-calibrated meta-analysis p values, even when the case-control ratios are extremely unbalanced and the minor allele counts are small. Through extensive numerical studies and an application on the UK Biobank data, we showed that the Z-score-based method can result in conservative or anti-conservative behavior in the meta-analysis p values, whereas our

proposed methods provided well-controlled type I error rates. The proposed methods also showed similar empirical powers as a joint analysis on the pooled data.

When the effect sizes are not available, such as in the case of the saddlepoint approximation-based test, it is widely popular to use the Z-score-based meta-analysis approach and combine the individual p values into a meta-analysis p value. In our third simulation setting, we showed that the Z-score-based approach can still be appropriate when only a small number of biobank-based studies with unbalance phenotypes are included in the meta-analysis. However, we will suggest the researchers to be cautious when using the Z-score-based approach, as including more such unbalanced studies can result in a loss of calibration in the meta-analysis p values. When effect size estimates are available, for example when using the Firth's bias-corrected likelihood ratio test (*Firth*, 1993), the inverse variance-weighted method is another popular meta-analysis approach used by the researchers. However, *Ma et al.* (2013) showed that the inverse variance-weighted meta-analysis method using the Firth's bias-corrected effect size estimates also results in type I error inflation when meta-analyzing several unbalanced studies.

In this chapter, we assumed that the individual studies do not have genetically related samples. In presence of related samples, the SAIGE test (*Zhou et al.*, 2018) can properly account for the sample relatedness and provide accurate p values in single studies with unbalanced case-control ratios. As the SAIGE p values are calculated using the saddlepoint approximation method based on the score statistic and its CGF, the spline-based meta-analysis method can still be applicable for combining multiple studies that are analyzed using SAIGE. However, the genotype count-based method may not be appropriate in such scenarios as the genotype-only model does not contain any information about the sample relatedness. The applicability of our methods in studies containing genetically related samples, is left for future research.

Comparing the two proposed methods, the spline-based method (CGF-Spline)

does not require any assumption on the effect of the non-genetic covariates since it reconstructs spline approximations of the null distributions of the score statistics and uses them to calculate the meta-analysis p values. Thus, it is more suitable to be applied regardless of the covariate effects. On the other hand, the genotype count-based method (GC) assumes relatively balanced non-genetic covariates with low covariate effects. However, the numerical simulations with very strong covariate effects (Figure I.4) also showed no sign of type I error inflation or deflation for this method. Another difference between the proposed methods is in their applicability on imputed dosage data. As the GC method requires the overall genotype counts to construct the genotype-only model, it is more suitable to be applied when the within-study analyses are performed on the best-called genotypes instead of dosages. The CGF-spline method is robust in this aspect as it can utilize the CGFs of the test statistics regardless of whether they were calculated from genotype or dosage data. However, in our UK Biobank data analysis example (Figure I.3), both our proposed methods showed no sign of inflation or deflation of type I errors, even when the within-study tests were performed on dosage data. Therefore, for practical application purposes, the genotype count-based method can be used to obtain accurate meta-analysis p values. One advantage of the genotype count-based method is that it is software-independent, and requires information which are more readily available compared to the spline-based method.

The proposed meta-analysis methods can be hybridized based on the availability of the summary level information. For example, suppose one study only provides the p value and direction of association, a second study additionally provides the genotype counts or minor allele count (if it is a rare variant), and a third study provides the score statistic and spline-based information. Then, a hybrid meta-analysis approach will be to use a normal reference distribution for the p value from the first study, and a reference distribution based on the genotype-only model for the p value from the

second study to calculate the converted scores and their corresponding CGFs. The CGF of the score statistic in the third study can be reconstructed based on spline approximation. Then, the final meta-analysis score will be the sum of those individual scores, and the corresponding CGF will be the sum of those individual CGFs. The meta-analysis p value can then be obtained using the saddlepoint approximation method. We implemented this hybrid meta-analysis approach along with all our proposed methods and the Z-score-based method in our R package **SPAtest** (available on CRAN).

CHAPTER VI

Conclusion

In this dissertation, we addressed some of the most challenging problems of analyzing large-scale biological datasets. To address the theoretical challenges due to the high dimensionality of the data, we discussed the asymptotic behaviors of PCA and PLS in high-dimensional regimes. In Chapter II, we derived consistent estimators of the population eigenvalues, angles between the sample and population eigenvectors, correlations between the sample and population PC scores, and developed methods to adjust the bias in the predicted PC scores. In Chapter III, we developed a two-stage PLS method to address the over-fitting and shrinkage problems of PLS regression in models with high-dimensional predictors. Next, we focused on some of the computational and methodological challenges of analyzing large-scale GWASs and PheWASs. In Chapter IV, we proposed a fast and accurate single-variant test, that is scalable to be applied for testing millions of variants across thousands of phenotypes in a typical EHR-based PheWAS. Our proposed test, fastSPA, is robust to handle extreme case-control imbalances and rare allele counts. We further developed two robust meta-analysis methods to efficiently and accurately combine association results from unbalanced case-control studies across multiple biobanks, in Chapter V. Our research in the high-dimensional methods provides the researchers with statistical tools for data visualization, confounder adjustment, and proper interpretation, modeling

and prediction in high-dimensional regimes. On the other hand, the methods developed for large-scale GWASs and PheWASs facilitates the researchers in the discovery of novel genotype-phenotype associations at a genome-wide or phenome-wide level, in reasonable computation time.

In this era of data explosion, the problems of high-dimensionality and computational scalability are extremely relevant, and the future generation of methodological research needs to focus on these problems. We dealt with some of these problems in this dissertation, but the scope for future research is vast in this domain. For example, in high-dimensional methods, other asymptotic regimes such as the ultra high-dimensional regime (*Lee et al.*, 2014b) and others (*Jung and Marron*, 2009) need to be explored. In unbalanced case-control GWASs, scalable gene-based tests, and adjusting for within-study sample relatedness in meta-analysis methods, are also important and immediate future research directions. The problems are certainly not limited to the ones mentioned here. As we keep on generating increasingly large amounts of data, new kinds of problems will emerge, and new kinds of solutions will be required. The goal is not to get overwhelmed by data, and this dissertation provided an important stepping stone towards that goal.

APPENDICES

APPENDIX A

Proof of Theorems 2.1, 2.2, 2.3 and 2.4

Proof of Theorem 2.1. The first part of the proof follows directly from Result 2.1 along with the fact that on the domain of the distant spikes, the ψ function is strictly increasing, and hence is left invertible. Since $\psi''(\alpha) > 0$ for any $\alpha > \sup \Gamma_H$, $\psi'(\alpha)$ is a strictly increasing function for $\alpha > \sup \Gamma_H$. Let $S_\psi > \sup \Gamma_H$ be a solution for $\psi'(\alpha) = 0$. Then for any $\alpha > \sup \Gamma_H$, $\psi'(\alpha) > 0$ if and only if $\alpha > S_\psi$. Therefore the interval (S_ψ, ∞) is the domain of the distance spikes, and ψ is a strictly increasing function on this interval. The second part follows from Lemma 1.2. □

Proof of Theorem 2.2. The proof closely follows the proof of Theorem 2 in *Mestre* (2008a). However, contrary to *Mestre* (2008a), we do not assume that the population LSD contains the generalized spikes. Thus, some of the derivation steps and results are substantially different from *Mestre* (2008a). We start the derivation by first noting that the quadratic forms $\hat{\eta}_k$ can be expressed as contour integrals of a special class of Stieltjes transforms of the sample covariance matrix. Let us define,

$$\hat{m}_p(z) := s_1^T \left(S_p - zI_p\right)^{-1} s_2 = \sum_{j=1}^{p} \frac{s_1^T e_j e_j^T s_2}{d_j - z}; \quad \forall z \in \mathbb{C}^+$$

where s_1 and s_2 are non-random vectors with uniformly bounded norms. *Girko* (1996) and *Mestre* (2006) showed that under the assumption that the population LSD contains the generalized spikes,

$$|\hat{m}_p(z) - m_p(z)| \xrightarrow{a.s.} 0; \quad \forall z \in \mathbb{C}^+ \tag{A.1}$$

where

$$m_p(z) = s_1^T \left[w(z)\Sigma_p - zI_p\right]^{-1} s_2 = \sum_{j=1}^{p} \frac{s_1^T E_j E_j^T s_2}{w(z)\lambda_j - z}.$$

The function $w(z)$ is defined as $w(z) = 1-\gamma-\gamma z b_F(z)$ where $b_F(z) = \int (\tau - z)^{-1} dF(\tau)$ is the Stieltjes transform of the sample LSD. It is easy to check, by the same arguments provided in *Mestre* (2006), that the result still holds when the generalized spikes are considered lying outside the support of the population LSD. The functions $\hat{m}_p, m_p$ and b_F can be extended to $\mathbb{C}^- = \{z \in \mathbb{C} : Im(z) < 0\}$ by defining $\hat{m}_p(z) = \hat{m}_p^*(z^*), m_p(z) = m_p^*(z^*)$ and $b_F(z) = b_F^*(z^*)$ for $z \in \mathbb{C}^-$ where z^* is the complex conjugate of z. With this definition, $|\hat{m}_p(z) - m_p(z)| \xrightarrow{a.s.} 0$ even when $z \in \mathbb{C}^-$. Now $\hat{\eta}_k$ can be expressed as an integral of $\hat{m}_p$,

$$\hat{\eta}_k = \frac{1}{2\pi i} \oint_{\partial \hat{\mathbb{R}}_y^-(k)} \hat{m}_p(z) dz,$$

where $i = \sqrt{-1}, y > 0$ and $\partial \hat{\mathbb{R}}_y^-(k)$ is the negatively (clockwise) oriented boundary of the rectangle $\hat{\mathbb{R}}_y(k) = \{z \in \mathbb{C} : \hat{a}_1 \leq Re(z) \leq \hat{a}_2, |Im(z)| \leq y\}$. $\hat{a}_1$ and $\hat{a}_2$ can be arbitrarily chosen provided that $\hat{\mathbb{R}}_y(k)$ contains only the sample eigenvalue d_k and no other sample eigenvalue. Then the following lemma gives the almost sure limit of $\hat{\eta}_k$.

Lemma 1.1.

$$\left| \frac{1}{2\pi i} \oint_{\partial \hat{\mathbb{R}}_y^-(k)} \hat{m}_p(z) dz - \frac{1}{2\pi i} \oint_{\partial \mathbb{R}_y^-(k)} m_p(z) dz \right| \xrightarrow{a.s.} 0,$$

where $y > 0$ and $\partial\mathbb{R}_y^-(k)$ is the negatively (clockwise) oriented boundary of the rectangle $\mathbb{R}_y(k) = \{z \in \mathbb{C} : a_1 \leq Re(z) \leq a_2, |Im(z)| \leq y\}$. a_1 and a_2 can be arbitrarily chosen so that $\psi(\lambda_k) \in [a_1, a_2]$ and $[a_1, a_2] \subset \psi(S_\psi, \infty)$ where $S_\psi > \sup \Gamma_H, \psi'(S_\psi) = 0$. $\psi(S_\psi, \infty)$ denotes the image of the interval (S_ψ, ∞) under ψ.

Lemma 1.1 implies

$$\left| \hat{\eta}_k - \sum_{j=1}^{p} \left(\frac{1}{2\pi i} \oint_{\partial\mathbb{R}_y^-(k)} \frac{dz}{w(z)\lambda_j - z} \right) s_1^T E_j E_j^T s_2 \right| \xrightarrow{a.s.} 0. \tag{A.2}$$

Now we need to evaluate the integral in (A.2) in order to get the almost sure limit of the random variable $\hat{\eta}_k$. First, we extend the ψ function to $\mathbb{R}_y(k)$ as follows,

$$\psi(z) := z \left(1 + \gamma \int \frac{\lambda dH(\lambda)}{z - \lambda} \right), \quad \forall z \in \mathbb{R}_y(k).$$

According to *Marčenko and Pastur* (1967), for all $z \in \mathbb{C}^+$, $b_F(z) = b$ is the unique solution to the following equation

$$b = \int \frac{dH(\lambda)}{\lambda(1 - \gamma - \gamma z b) - z} \tag{A.3}$$

in the set $\{b \in \mathbb{C} : \gamma b - (1 - \gamma)/z \in \mathbb{C}^+\}$. It is easy to see that b_F also satisfies (A.3) when $z \in \mathbb{C}^-$. Now we formally define the f_F function introduced in (2.1),

$$f_F(z) := \frac{z}{w(z)} = \frac{z}{1 - \gamma - \gamma z b_F(z)}, \quad \forall z \in \mathbb{C} \setminus \mathbb{R}. \tag{A.4}$$

Then b_F can be expressed in terms of f_F as,

$$b_F(z) = \frac{(1 - \gamma) f_F(z) - z}{\gamma z f_F(z)}.$$

By replacing b with $[(1-\gamma)f - z]/\gamma z f$ in (A.3),

$$f\left(1+\gamma\int\frac{\lambda dH(\lambda)}{f-\lambda}\right)=z. \tag{A.5}$$

It is easy to see that b_F is a solution to (A.3) if and only if f_F is a solution to (A.5). Therefore, for all $z\in\mathbb{C}^+$ (similarly for $z\in\mathbb{C}^-$), $f_F(z)=f$ is the unique solution to (A.5) on $\mathbb{C}^+$ (respectively, $\mathbb{C}^-$). This implies $\psi\left(f_F(z)\right)=z$ for all $z\in\mathbb{R}_y(k)\setminus[a_1,a_2]$.

Now we focus on the case when $z\in\mathbb{R}\backslash\{0\}$. According to *Silverstein and Choi* (1995), we can extend b_F to $\mathbb{R}\backslash\{0\}$ by defining $b_F(z)=\lim_{y\to 0^+}b_F(z+iy)$ for any $z\in\mathbb{R}\backslash\{0\}$. The definition of f_F can also be extended in a similar fashion. In Lemma 1.2 we have shown that f_F is the inverse function of ψ on (S_ψ,∞), and there exists $M_f>\sup\Gamma_F$ for which $\psi\left(S_\psi,\infty\right)=(M_f,\infty)$. Thus, $[a_1,a_2]\subset\psi\left(S_\psi,\infty\right)$ implies $\psi\left(f_F(z)\right)=z$ for all $z\in\mathbb{R}_y(k)$. Furthermore, the function ψ is continuous and differentiable on $\mathbb{R}_y(k)$, and the derivative is given by,

$$\psi'(z)=1-\gamma\int\left(\frac{\lambda}{z-\lambda}\right)^2 dH(\lambda).$$

Then the integral in (A.2) can be expressed in terms of ψ and f_F as follows,

$$\begin{aligned}\frac{1}{2\pi i}\oint\limits_{\partial\mathbb{R}_y^-(k)}\frac{dz}{w(z)\lambda_j-z}&=\frac{1}{2\pi i}\oint\limits_{\partial\mathbb{R}_y^-(k)}\frac{dz}{\frac{z}{f_F(z)}\lambda_j-z}\\&=\frac{1}{2\pi i}\oint\limits_{\partial\mathbb{R}_y^-(k)}\frac{1}{\lambda_j-f_F(z)}.\frac{f_F(z)}{\psi\left(f_F(z)\right)}dz.\end{aligned} \tag{A.6}$$

The integrand in the final expression is holomorphic on $\mathbb{R}_y^-(k)$ when $j\neq k$ and possesses a simple pole $\psi(\lambda_k)$ when $j=k$. Therefore, when $j\neq k$ the integral in

(A.6) is zero. When $j = k$, applying the residue theorem on the final integral,

$$\begin{aligned}\frac{1}{2\pi i}\oint_{\partial\mathbb{R}_y^-(k)} \frac{dz}{w(z)\lambda_k - z} &= \lim_{z\to\psi(\lambda_k)} \frac{\psi(\lambda_k) - z}{\lambda_k - f_F(z)} . \frac{f_F(z)}{\psi\left(f_F(z)\right)} \\ &= \lim_{z\to\psi(\lambda_k)} \frac{\psi(\lambda_k) - \psi\left(f_F(z)\right)}{\lambda_k - f_F(z)} . \frac{f_F(z)}{\psi\left(f_F(z)\right)} \\ &= \frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)}.\end{aligned}$$

This implies,

$$\frac{1}{2\pi i}\oint_{\partial\mathbb{R}_y^-(k)} \frac{dz}{w(z)\lambda_j - z} = \begin{cases} \frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)} & j = k \\ 0 & j \neq k \end{cases}$$

and the proof is complete. □

Proof of Lemma 1.1. First, we show that $\hat{a}_1, \hat{a}_2, a_1, a_2$ can be chosen satisfying $\hat{a}_1 \to a_1$ and $\hat{a}_2 \to a_2$. This is possible due to the fact that $d_k \xrightarrow{a.s.} \psi(\lambda_k)$ and $\psi(\lambda_k) \subset \psi\left(S_\psi, \infty\right) = (M_f, \infty)$ where $M_f > \sup \Gamma_F$. Therefore, we can choose a neighborhood $[a_1, a_2]$ around $\psi(\lambda_k)$ so that $[a_1, a_2] \subset (M_f, \infty)$. Moreover, as M_f is bounded away from the support of the sample LSD F and $d_k \xrightarrow{a.s.} \psi(\lambda_k)$, we can select a neighborhood $[\hat{a}_1, \hat{a}_2]$ around d_k which does not contain any other eigenvalue for which $\hat{a}_1 \to a_1, \hat{a}_2 \to a_2$. Then,

$$\begin{aligned}&\left| \frac{1}{2\pi i}\oint_{\partial\hat{\mathbb{R}}_y^-(k)} \hat{m}_p(z)dz - \frac{1}{2\pi i}\oint_{\partial\mathbb{R}_y^-(k)} m_p(z)dz \right| \\ &\qquad \leq \frac{1}{2\pi}\left\{ \sup_{z\in\partial\hat{\mathbb{R}}_y^-(k)\cup\partial\mathbb{R}_y^+(k)} |\hat{m}_p(z)| \right\} \left(|\hat{a}_1 - a_1)| + |\hat{a}_2 - a_2)|\right) \\ &\qquad\qquad + \frac{1}{2\pi}\oint_{\partial\mathbb{R}_y^-(k)} |\hat{m}_p(z)dz - m_p(z)|\,|dz|. \quad \text{(A.7)}\end{aligned}$$

From Cauchy-Schwartz inequality, we can obtain the following upper bound for $\hat{m}_p$,

$$|\hat{m}_p(z)| \leq \frac{\|s_1\|\|s_2\|}{d\left(z, \Gamma_{F_p}\right)},$$

where $d\left(z, \Gamma_{F_p}\right) = \inf_{y \in \Gamma_{F_p}} |z - y|$. Since $F_p \to F$ point-wise and $[a_1, a_2]$ is bounded away from Γ_F, $d\left(z, \Gamma_{F_p}\right)$ is bounded away from zero with probability one for large enough p and n. Therefore $|\hat{m}_p(z)|$ is finite for $z \in \mathbb{R}_y(k)$ with probability one for large enough p and n. Moreover, since $[\hat{a}_1, \hat{a}_2] \to [a_1, a_2]$, the interval $[\hat{a}_1, \hat{a}_2]$ will eventually be bounded away from Γ_F. Thus, eventually the upper bound for $|\hat{m}_p(z)|$ will also be finite for $z \in \hat{\mathbb{R}}_y(k)$. Therefore, the first term on the right hand side of (A.7) will go to zero as $\hat{a}_1 \to a_1, \hat{a}_2 \to a_2$.

Now, as $\hat{m}_p(z)$ and $m_p(z)$ are holomorphic functions on the compact set $\partial\mathbb{R}_y^-(k)$,

$$\sup_{z \in \partial\mathbb{R}_y^-(k)} |\hat{m}_p(z) - m_p(z)| < \infty.$$

Also from (A.1), $|\hat{m}_p(z) - m_p(z)| \xrightarrow{a.s.} 0$ point-wise for all $z \in \mathbb{C} \setminus \mathbb{R}$. Therefore, by dominated convergence theorem the second term on the right hand side of (A.7) also converges to zero almost surely. □

We can show the asymptotic equivalence of the limits derived in Theorem 2.2 and Result 2.2 as a direct application of the following lemma.

Lemma 1.2. *Suppose Assumptions 2.1–2.3 hold. If λ_k is a distant spike with multiplicity one, and d_k is the corresponding sample eigenvalue, then*

$$f_F(d_k) \xrightarrow{p} \lambda_k; \quad \frac{d_k g_F(d_k)}{f_F(d_k)} \xrightarrow{p} \psi'(\lambda_k).$$

Proof. We have already established in the proof of Theorem 2.2 that for all $z \in \mathbb{C}^+$ (similarly for $z \in \mathbb{C}^-$), $f_F(z) = f$ is the unique solution to (A.5) on $\mathbb{C}^+$ (respectively, $\mathbb{C}^-$). When z is restricted to $\mathbb{C} \setminus \mathbb{R}$, using (A.4) and the fact that $b_F(z) =$

$\int (\tau - z)^{-1} dF(\tau)$ we can write,

$$f_F(z) = \frac{z}{1 + \gamma \int \frac{\tau dF(\tau)}{z - \tau}}.$$

Now suppose $z = x \in \mathbb{R} \setminus \{0\}$. Then both equations (A.3) and (A.5) will have multiple roots (both real and complex valued depending on x and H). If we look at (A.5) closely, we can see for real valued x it can be represented as $\psi(f(x)) = x$, where the ψ function is as defined in (2.1). As we have seen in the proof of Theorem 2.1, ψ is strictly increasing in the interval (S_ψ, ∞) where $S_\psi > \sup \Gamma_H$ and $\psi'(S_\psi) = 0$. Therefore, any real-valued solution f of $\psi(f(x)) = x$ in (S_ψ, ∞) has to be the inverse of ψ, which is unique due to the strict monotonicity of ψ on (S_ψ, ∞). Now suppose Γ_F is the support of the sample LSD F. We will show that there exists $M_f > \sup \Gamma_F$ such that for any $x > M_f$, the function f_F is real-valued and it is a solution to (A.5) in the interval (S_ψ, ∞). Thus it is also the unique such solution and the inverse of the ψ function in (S_ψ, ∞).

Let $x \in \mathbb{R}, x > \sup \Gamma_F$ and $z = x + iy \in \mathbb{C}^+$. Now, as $z \in \mathbb{C}^+$, $f_F(z)$ is the unique solution to (A.5) in $\mathbb{C}^+$. Therefore, if we express $f_F(z)$ as $u(z) + iv(z)$, then $v(z) > 0$. Also, the imaginary part of (A.5) can be written as

$$v(z) \left[1 - \gamma \int \frac{\lambda^2 dH(\lambda)}{\{u(z) - \lambda\}^2 + v(z)^2} \right] = y.$$

Both $v(z)$ and y being positive implies that

$$1 - \gamma \int \frac{\lambda^2 dH(\lambda)}{\{u(z) - \lambda\}^2 + v(z)^2} > 0. \tag{A.8}$$

Due to the continuity of f_F on the set $\{z \in \mathbb{C}^+ : z = x + iy, x > \sup \Gamma_F\}$,

$$f_F(x) = \lim_{y \to 0^+} \frac{x + iy}{1 + \gamma \int \frac{\tau dF(\tau)}{x + iy - \tau}} = \frac{x}{1 + \gamma \int \frac{\tau dF(\tau)}{x - \tau}},$$

which is real-valued. Thus $u(z) \to f_F(x)$ and $v(z) \to 0$ as $y \to 0^+$. Therefore as $y \to 0^+$, the inequality (A.8) becomes

$$1 - \gamma \int \frac{\lambda^2 dH(\lambda)}{\{f_F(x) - \lambda\}^2} > 0,$$

which implies $\psi'(f_F(x)) > 0$.

We can see that $f_F(x)$ attains zero at $\sup \Gamma_F$ and it is strictly and unboundedly increasing for $x > \sup \Gamma_F$. This ensures the existence of a threshold $M_F > \sup \Gamma_F$ such that the function f_F maps the interval (M_F, ∞) to (S_ψ, ∞). Therefore, f_F and ψ are both strictly increasing, continuous and bijective mappings between the intervals (M_F, ∞) and (S_ψ, ∞). Since $f_F(z)$ is the unique solution to (A.5) in $\mathbb{C}^+$ when $z \in \mathbb{C}^+$, f_F is also a solution to (A.5) in (S_ψ, ∞) when $x > M_F$ due to the continuity of the left hand side of (A.5) on the set $\{f \in \mathbb{C}^+ : f = u + iv, u > S_\psi\}$, which further implies that f_F is the inverse function of ψ on (S_ψ, ∞).

The first part of this lemma is proved as a corollary to Result 2.1 as $\psi^{-1} = f_F$ on the domain of distant spikes, i.e. (S_ψ, ∞). For the second part we first need to derive the expression of f_F', and then derive the expression of ψ' in terms of f_F and F.

$$f_F'(x) = \frac{f(x)}{x}\left[1 + \gamma v_F(x)\right]; \quad v_F(x) = \int \frac{\tau dF(\tau)}{(x - \tau)^2}.$$

For a distant spike λ_k, using the expression of f_F' we get,

$$\frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)} = \frac{\lambda_k}{\psi(\lambda_k) f_F'(\psi(\lambda_k))} = \frac{1}{1 + \gamma f_F(\psi(\lambda_k)) \int \frac{\tau dF(\tau)}{[\psi(\lambda_k) - \tau]^2}} = g_F(\psi(\lambda_k)).$$

As $\psi(\lambda_k) > M_f$, g_F is continuous at $\psi(\lambda)$. Since $d_k \xrightarrow{p} \psi(\lambda_k)$,

$$g_F(d_k) \xrightarrow{p} g_F(\psi(\lambda_k)) = \frac{\lambda_k \psi'(\lambda_k)}{\psi(\lambda_k)}; \quad \frac{d_k g_F(d_k)}{f_F(d_k)} \xrightarrow{p} \psi'(\lambda_k).$$

□

Proof of Theorem 2.3.

$$\langle P_k, p_k\rangle^2 = \frac{1}{n^2\lambda_k d_k}\langle XE_k, Xe_k\rangle^2 = \frac{1}{\lambda_k d_k}\left(E_k^T\frac{X^TX}{n}e_k\right)^2$$
$$= \frac{1}{\lambda_k d_k}\left[E_k^T\left(\sum_{i=1}^{p} d_i e_i e_i^T\right)e_k\right]^2 = \frac{d_k}{\lambda_k}\langle e_k, E_k\rangle^2 .$$

Using the limits derived in Theorem 2.2 and Result 2.1,

$$\left|\frac{d_k}{\lambda_k}\langle e_k, E_k\rangle^2 - \psi'(\lambda_k)\right| \xrightarrow{p} 0.$$

Using Lemma 1.2,

$$\left|\frac{d_k}{\lambda_k}\langle e_k, E_k\rangle^2 - \frac{d_k g_F(d_k)}{f_F(d_k)}\right| \xrightarrow{p} 0.$$

□

Proof of Theorem 2.4. We show that the denominator $E\left(p_{kj}^2\right)$ converges to $\psi(\lambda_k)$ and the numerator $E\left(q_k^2\right)$ converges to $\lambda_k^2/\psi(\lambda_k)$. The proof will be complete using the fact that $d_k \xrightarrow{p} \psi(\lambda_k)$.

The denominator,

$$E\left(p_{kj}^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} p_{ki}^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}(x_i^Te_k)^2\right)$$
$$= E\left(e_k^T\frac{X^TX}{n}e_k\right) = E\left[e_k^T\left(\sum_{i=1}^{p} d_i e_i e_i^T\right)e_k\right] = E(d_k) \to \psi(\lambda_k).$$

The numerator,

$$E\left(q_k^2\right) = E\left[(x_{new}^Te_k)^2\right] = E\left[E(x_{new}^Te_k)^2|e_k\right]$$
$$= E\left[Var(x_{new}^Te_k)|e_k\right] = E\left[e_k^T\Sigma_p e_k\right].$$

Now, using the notations in the proof of Theorem 2.2 and Lemma 1.2, we have

$b_F(z) = \int (\tau - z)^{-1} dF(\tau)$ as the Stieltjes transform of the sample LSD and the function f_F defined as $f_F(z) = z\left[1 - \gamma - \gamma z b_F(z)\right]^{-1}$. Therefore,

$$b_F(z) = \frac{(1-\gamma) f_F(z) - z}{\gamma z f_F(z)}.$$

The functions b_F and f_F can be extended to the real axis by defining the extensions as shown in the proof of Lemma 1.2. Thus, for the sample eigenvalue d_k corresponding to the distant spike λ_k we have

$$b_F(d_k) = \frac{(1-\gamma) f_F(d_k) - d_k}{\gamma d_k f_F(d_k)}.$$

According to Theorem 4 in *Ledoit and Péché* (2010), the limit of $e_k^T \Sigma_p e_k$ is given by $d_k\left[1 - \gamma - \gamma d_k b_F(d_k)\right]^{-2}$. Replacing the expression of $b_F(d_k)$ in this limit, we get

$$\left| e_k^T \Sigma_p e_k - \frac{f_F^2(d_k)}{d_k} \right| \xrightarrow{p} 0.$$

Using Result 2.1 and Lemma 1.2 we have $f_F^2(d_k)/d_k \xrightarrow{p} \lambda_k^2/\psi(\lambda_k)$. Therefore, the limit of the numerator is given by,

$$E(q_k^2) = E\left[e_k^T \Sigma_p e_k\right] \to \frac{\lambda_k^2}{\psi(\lambda_k)}.$$

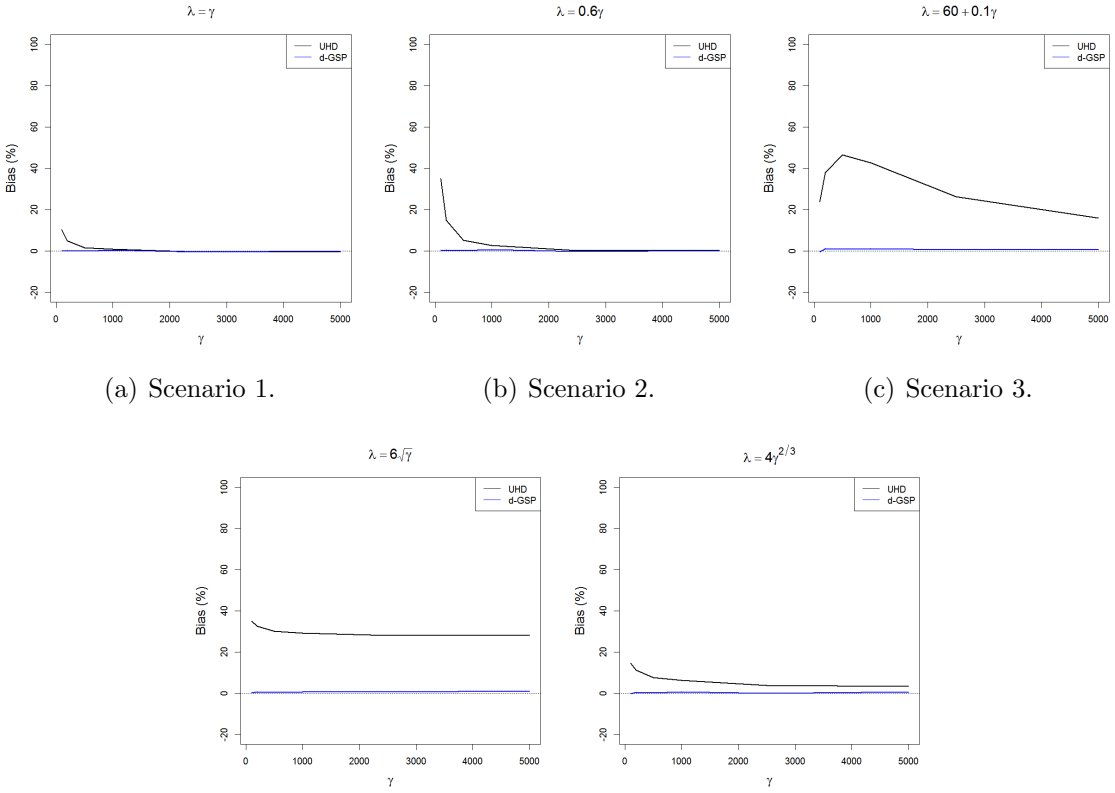
□

APPENDIX B

Supplementary Tables and Figures for Chapter II

Settings					**Estimated no. of distant spikes**			
No.	n	p	σ^2	ρ	1	2	3	≥ 4
1	500	5000	4	0.8	0	38.5	38.5	23
2	500	5000	1	0.7	0	69.5	26.5	4
3	500	5000	7.5	0.8	47.5	39	10	3.5
4	500	5000	4	0	0	94	6	0

Table B.1: Percentage of simulated datasets where the number of distant spikes were estimated to be $1, 2, 3$ or ≥ 4.



(a) Scenario 1. (b) Scenario 2. (c) Scenario 3.

(d) Scenario 4. (e) Scenario 5.

Figure B.1: Empirical biases (%) in estimating the shrinkage factor corresponding to the largest population eigenvalue for GSP-based and UHD-based methods. The population eigenvalues and the rate of increment of the largest population eigenvalue are assumed to be unknown.

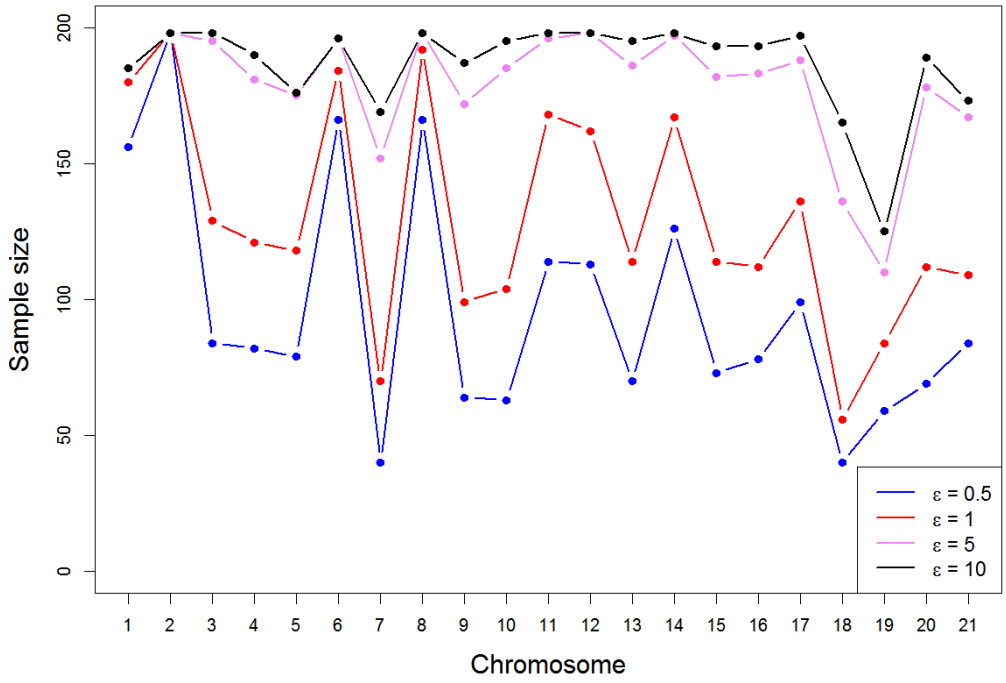


Figure B.2: Sample sizes of the test samples that were included in the prediction error estimation for different values of the thresholding parameter ϵ

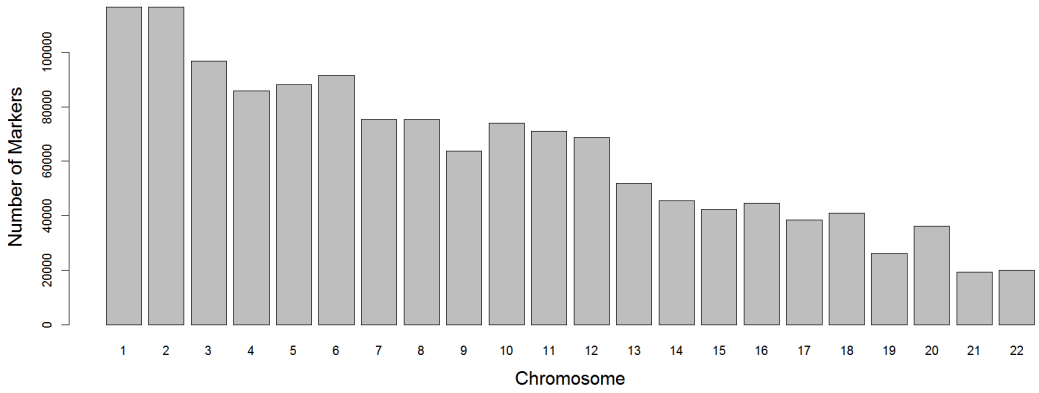


Figure B.3: Distribution of the number of markers across different chromosomes.

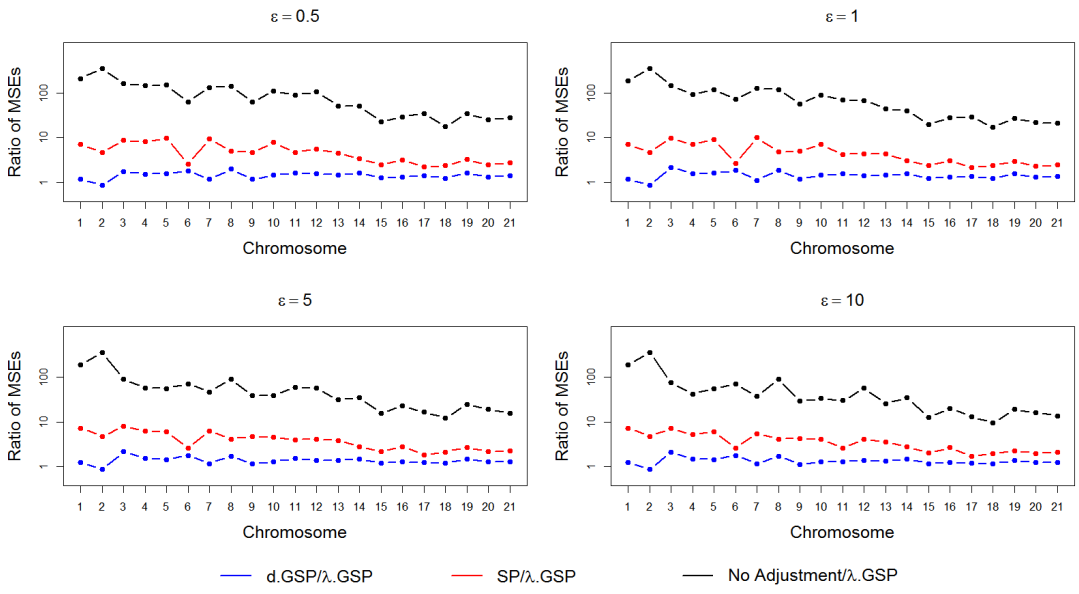


Figure B.4: Comparison of the mean squared errors (MSE) of the unadjusted, d-GSP-adjusted, and SP-adjusted PC scores, with the λ-GSP-adjusted PC scores using different values of the thresholding parameter ϵ. The ratios of the MSEs are presented for chromosome 1-21. The Y-Axis is presented in a logarithmic scale.

APPENDIX C

Proof of Theorems 3.1 and 3.2

Proof of Theorem 3.1. Let $\Sigma_p = U\Lambda U^T$ be the eigen-decomposition of Σ_p. As $rank(T) = k$, the first k eigenvalues $\lambda_1, \ldots, \lambda_k$ of Σ_p will be non-zero, and the rest will be zero. Let $Q^T = P^T R$ (such R exists as $rank(P^T) = k$), $U^T R = A = [a_1 \quad a_2 \quad \ldots \quad a_q]$. Further, for any $p \times 1$ orthonormal vector w, assume $\gamma = (\gamma_1, \ldots, \gamma_p) = U^T w$. $w^T w = 1$ if and only if $\gamma^T \gamma = 1$. Then,

$$\begin{aligned} n^{-2} w^T P T^T T Q^T Q T^T T P^T w &= w^T \Sigma_p R R^T \Sigma_p w \\ &= \gamma U^T \Sigma_p U A A^T U^T \Sigma_p U \gamma \\ &= \gamma \Lambda A A^T \Lambda \gamma \\ &= \sum_{j=1}^{q} \left(a_j^T \Lambda \gamma \right)^2 \\ &= \sum_{j=1}^{q} \left(\sum_{l=1}^{k} \lambda_l a_{jl} \gamma_l \right)^2 \end{aligned}$$

Therefore, $w = w_i$ maximizes $n^{-2} w^T P T^T Y Y^T T P^T w$ with constraints $w^T w = 1$ and $w^T S_{xx} w_j = 0$ for $j = 1, \ldots, i-1$, if and only if $\gamma = U^T w_i$ maximizes $\sum_{j=1}^{q} \left(\sum_{i=l}^{k} \lambda_l a_{jl} \gamma_l \right)^2$ with constraints $\gamma^T \gamma = 1$ and $\gamma^T U^T S_{xx} w_j = 0$ for all $j = 1, \ldots, i-1$. The proof is complete by noting that any γ (with the aforementioned constraints) which maximizes

$\sum_{j=1}^{q}\left(\sum_{l=1}^{k}\lambda_l a_{jl}\gamma_l\right)^2$ must have all $\gamma_{k+1}=\ldots=\gamma_p=0$ as the objective function does not involve those elements. □

Proof of Theorem 3.2. It is easy to verify that the eigenvalues and angles between the sample and population eigenvectors are rotation invariant. Therefore, without loss of generality, we assume that the population covariance matrix is diagonal, i.e., the population eigenvectors $u_1,\ldots,u_p$ are the p-dimensional euclidean basis vectors.

Now, if $1<\theta_i\leq 1+\sqrt{\gamma}$, i.e., the eigenvalue θ_i is a close spike, then the proof is given as a part of the proof for Lemma 2 in *Lee et al.* (2010). If we partition the sample eigenvector v_i as (v_{Ai}, v_{Bi}), where A represents the first k coordinates, and B represents the rest. If we define $R_i=||v_{Bi}||$, then *Lee et al.* (2010) showed that $R_i\xrightarrow{p}0$ for $1<\theta_i\leq 1+\sqrt{\gamma}$. As $R_i=\sqrt{\sum_{j=1}^{k}\langle v_i,u_j\rangle^2}$, each individual $\langle v_i,u_j\rangle\xrightarrow{p}0$.

Next, we consider the case when $\theta_i=1$ and $d_i>0$, i.e., the eigenvalue θ_i is a non-spike. To prove this, we use the theoretical results in *Ledoit and Péché* (2010), specifically Theorem 3. Following the definitions given in *Ledoit and Péché* (2010), we define the following function in our notation,

$$\Phi_p(d,\lambda)=\frac{1}{p}\sum_{r=1}^{p}\sum_{j=1}^{p}\langle v_r,u_j\rangle^2\,1_{d\geq d_i}(d)\times 1_{\lambda\geq\lambda_j}(\lambda)\quad\forall d,\lambda\in\mathbb{R}\tag{C.1}$$

Ledoit and Péché (2010) showed that Φ_p satisfies the properties of a bivariate distribution function. We assume the sample ESD and LSD to be F_p and F respectively, and population ESD and LSD to be H_p and H respectively. Notice that, for spiked population models, H is a degenerate distribution at unity. We further define the Stieltjes transform of F to be b, and $\tilde{b}(d)=\lim_{z\in\mathbb{C}^+\to d}b(z)$. Then,

$$\Phi(d,1)=\frac{1}{p}\sum_{r=1}^{p}\sum_{j=k+1}^{p}\langle v_r,u_j\rangle^2\,1_{d\geq d_i}(d),$$

which implies,

$$\sum_{j=k+1}^{p} \langle v_i, u_j \rangle^2 = \lim_{h \to 0} \frac{\Phi_p(d_i + h, 1) - \Phi_p(d_i - h, 1)}{F_p(d_i + h) - F_p(d_i - h)}.$$

Therefore, as $p \to \infty$,

$$\begin{aligned}
\sum_{j=k+1}^{p} \langle v_i, u_j \rangle^2 &\xrightarrow{a.s.} \lim_{h \to 0} \frac{\Phi(d_i + h, 1) - \Phi(d_i - h, 1)}{F(d_i + h) - F(d_i - h)} \\
&= \frac{\partial \Phi(d, 1)}{\partial F(d)}|_{d=d_i} \\
&= \phi(d_i, 1),
\end{aligned}$$

where Φ and ϕ are given by Theorem 3 in *Ledoit and Péché* (2010). When $\gamma > 1$, $\Phi_p(d, \lambda) \xrightarrow{a.s.} \Phi(d, \lambda)$ at all points of continuity of Φ, where $\Phi(d, \lambda) = \int_{-\infty}^{d} \int_{-\infty}^{\lambda} \phi(\delta, l) dH(l) dF(\delta)$, and

$$\phi(\delta, l) = \frac{\gamma \delta l}{(pl - \delta)^2 + q^2 l^2}, \quad \text{if } \delta > 0,$$

p and q are the real and imaginary parts of $1 - \gamma - \gamma \delta \tilde{b}(\delta)$, respectively. Therefore, if we show that $\phi(d_i, 1) = 1$, then the proof is complete, as $\sum_{j=1}^{k} \langle v_i, u_j \rangle^2 = 1 - \sum_{j=k+1}^{p} \langle v_i, u_j \rangle^2 \xrightarrow{a.s.} 0$.

Now, from the Marčenko–Pastur theorem (*Marčenko and Pastur*, 1967),

$$\begin{aligned}
b(z) &= \int \frac{dH(\lambda)}{\lambda(1 - \gamma - \gamma z b(z)) - z} \\
&= \frac{1}{1 - \gamma - \gamma z b(z) - z}.
\end{aligned} \tag{C.2}$$

Let us assume $z = d_i + is$ and $b(z) = x + iy$ where $x, y, s \in \mathbb{R}$. Then, solving (C.2),

$$\begin{aligned}
\frac{x}{x^2 + y^2} &= 1 - \gamma - \gamma d_i x + \gamma s y - d_i \\
\frac{y}{x^2 + y^2} &= \gamma s x + \gamma d_i y + s.
\end{aligned}$$

Letting $s \to 0$,

$$\frac{x}{x^2+y^2} \to 1-\gamma-\gamma d_i x - d_i$$
$$\frac{y}{x^2+y^2} \to \gamma d_i y \implies x^2+y^2 = \frac{1}{\gamma d_i}.$$

Solving for x and y, we get,

$$x \to \frac{1-\gamma-d_i}{2\gamma d_i}$$
$$y \to \frac{\sqrt{4\gamma d_i - (1-\gamma-d_i)^2}}{2\gamma d_i}.$$

Therefore, p and q, the real and imaginary parts of $1-\gamma-\gamma d_i \tilde{b}(d_i)$ are given by,

$$p = \frac{1-\gamma+d_i}{2}$$
$$q = \frac{\sqrt{4\gamma d_i - (1-\gamma-d_i)^2}}{2}.$$

Finally, replacing p and q into the expression for ϕ,

$$\begin{aligned}\phi(d_i, 1) &= \frac{\gamma d_i}{(p-d_i)^2+q^2} \\ &= 1\end{aligned}$$

and the proof is complete. □

APPENDIX D

Explanation Behind Using the Covariate-Adjusted Genotypes ($\tilde{G}$) in the Expression of the Score Statistic

We first note that $S = \tilde{G}^T (Y - \hat{\mu}) = G^T (Y - \hat{\mu})$ since $\hat{\mu}$ is the maximum likelihood estimator of μ under the null model and $X^T (Y - \hat{\mu}) = 0$. Now, the score function and the observed information matrix under the null model are given by,

$$U_0 = \begin{bmatrix} X^T (Y - \hat{\mu}) \\ G^T (Y - \hat{\mu}) \end{bmatrix} = \begin{bmatrix} 0 \\ S \end{bmatrix}, \quad I_0 = \begin{bmatrix} X^T W X & X^T W G \\ G^T W X & G^T W G \end{bmatrix}.$$

Therefore, the variance of S under H_0 is given by,

$$V_{H_0}(S) = G^T W G - G^T W X \left(X^T W X\right)^{-1} X^T W G = G^T W \tilde{G} = \tilde{G}^T W \tilde{G}.$$

So, even though the two expressions of S are algebraically equivalent, the variance can be expressed as a weighted sum of $\hat{\mu}_i (1 - \hat{\mu}_i)$s where the eights are given by $\tilde{G}_i$s. Therefore, we used $\tilde{G}$ instead of G to express the score statistic.

APPENDIX E

Supplementary Tables and Figures for Chapter IV

		Genomic Control at q^{th} p value quantile		
Case:Control	**Test**	$q = 0.5$ (Median)	$q = 0.01$	$q = 0.01$
10000:10000	fastSPA-2	1	1	1
	fastSPA-BE	1	1	1
	fastSPA-0.1	1	1	1
2000:18000	fastSPA-2	1.01	1	1
	fastSPA-BE	1	1	1
	fastSPA-0.1	1	1	1
40:19960	fastSPA-2	0.48	0.99	0.99
	fastSPA-BE	1.83	0.99	0.99
	fastSPA-0.1	1.83	0.99	0.99

Table E.1: Estimated inflation factors of the genomic controls at different p value quantiles for the fastSPA-2, fastSPA-BE and fastSPA-0.1 tests applied on 5×10^6 simulated variants. The significance level for the fastSPA-BE test was selected to be $\alpha = 5 \times 10^{-8}$. MAFs of variants were randomly sampled from the MAF distribution of the MGI data.

Phenotype	Test	MAF cutoff	Genomic Control at q^{th} p value quantile $q = 0.5$ (Median)	$q = 0.01$	$q = 0.001$
Skin Cancer	fastSPA-2	All variants	1.11	1.02	1.03
		> 0.001	1.02	1.02	1.04
		> 0.01	1.01	1.03	1.05
	fastSPA-BE	All variants	1.02	1.02	1.03
		> 0.001	1.02	1.02	1.04
		> 0.01	1.01	1.03	1.05
	fastSPA-0.1	All variants	1.01	1.02	1.03
		> 0.001	1.01	1.02	1.04
		> 0.01	1.01	1.03	1.05
Type-2 Diabetes	fastSPA-2	All variants	1.11	1.02	1.01
		> 0.001	1.02	1.02	1.02
		> 0.01	1.03	1.02	1.02
	fastSPA-BE	All variants	1.00	1.02	1.01
		> 0.001	1.02	1.02	1.02
		> 0.01	1.02	1.02	1.02
	fastSPA-0.1	All variants	1.00	1.02	1.01
		> 0.001	1.02	1.02	1.02
		> 0.01	1.02	1.02	1.02
Primary Hypercoagulable State	fastSPA-2	All variants	0.37	1.02	0.98
		> 0.001	1.03	0.99	1.01
		> 0.01	1.00	1.00	1.02
	fastSPA-BE	All variants	1.04	1.02	0.98
		> 0.001	0.96	0.99	1.01
		> 0.01	1.02	1.00	1.02
	fastSPA-0.1	All variants	1.04	1.02	0.98
		> 0.001	0.96	0.99	1.01
		> 0.01	1.01	1.00	1.02
Cystic Fibrosis	fastSPA-2	All variants	0.12	0.99	1.01
		> 0.001	0.62	0.98	0.98
		> 0.01	1.07	0.99	0.98
	fastSPA-BE	All variants	1.27	1.00	1.01
		> 0.001	0.93	0.98	0.98
		> 0.01	1.07	0.99	0.98
	fastSPA-0.1	All variants	1.27	1.00	1.01
		> 0.001	0.94	0.98	0.98
		> 0.01	1.07	0.99	0.98

Table E.2: Estimated inflation factor of the genomic controls at different p value quantiles and different MAF cut-offs for the fastSPA-2, fastSPA-BE and fastSPA-0.1 tests applied on four different phenotypes from the MGI data. The genome-wide significance level for the fastSPA-BE test was selected to be $\alpha = 5 \times 10^{-8}$. Only the imputed variants were removed when we used different MAF cutoffs. The SNPs present on the Illumina HumanCoreExome v12.1 array were preserved.

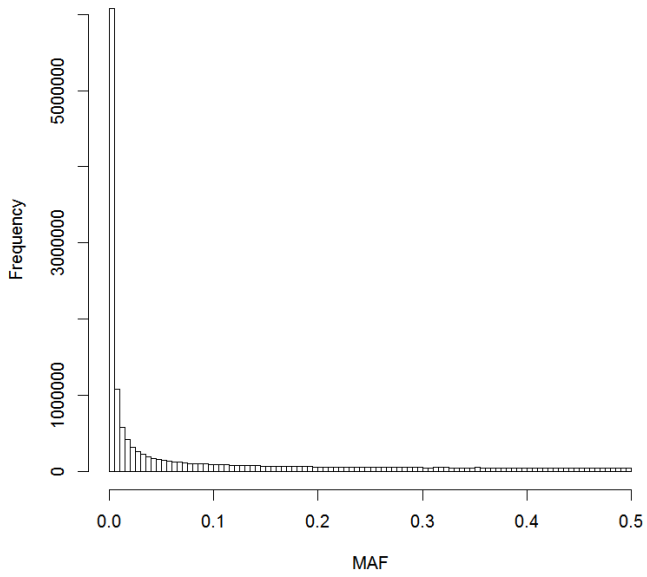


Figure E.1: Histogram of MAFs from the MGI data.

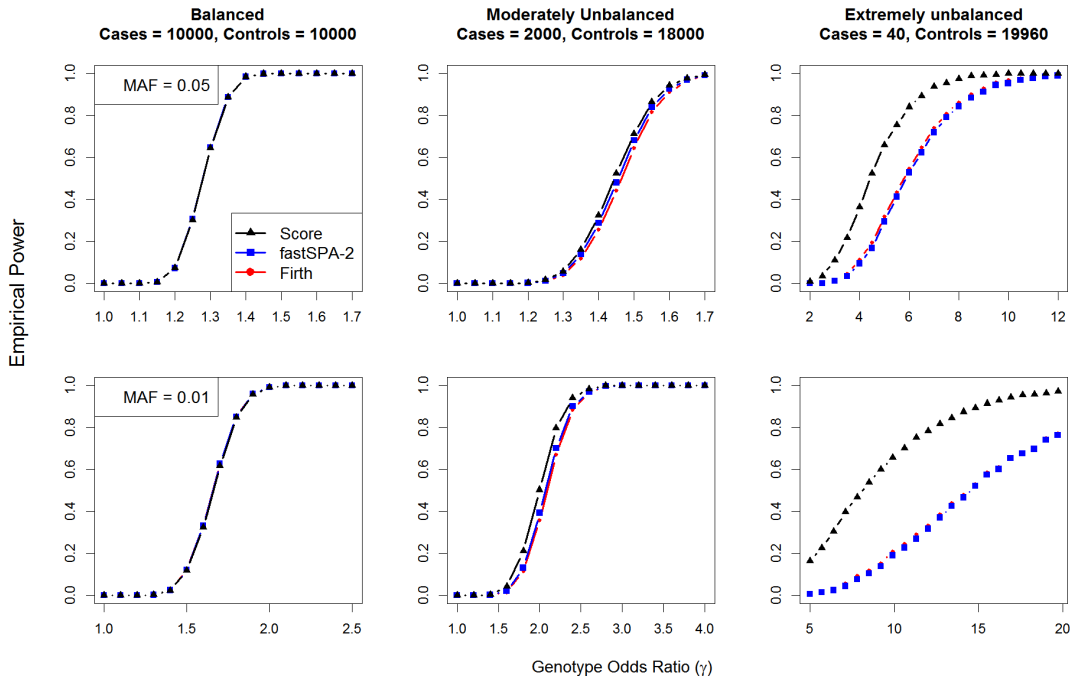


Figure E.2: Empirical power curves for the traditional score, fastSPA-2 and Firth tests at the nominal type I error level $\alpha = 5 \times 10^{-8}$ based on 5000 simulated datasets. Top panel considers MAF = 0.05 and bottom panel considers MAF = 0.01. From left to right, the plots consider case:control = 10000:10000, 2000:18000 and 40:19960, respectively. In each plot x-axis represents genotype odds ratios and y-axis represents the empirical power.

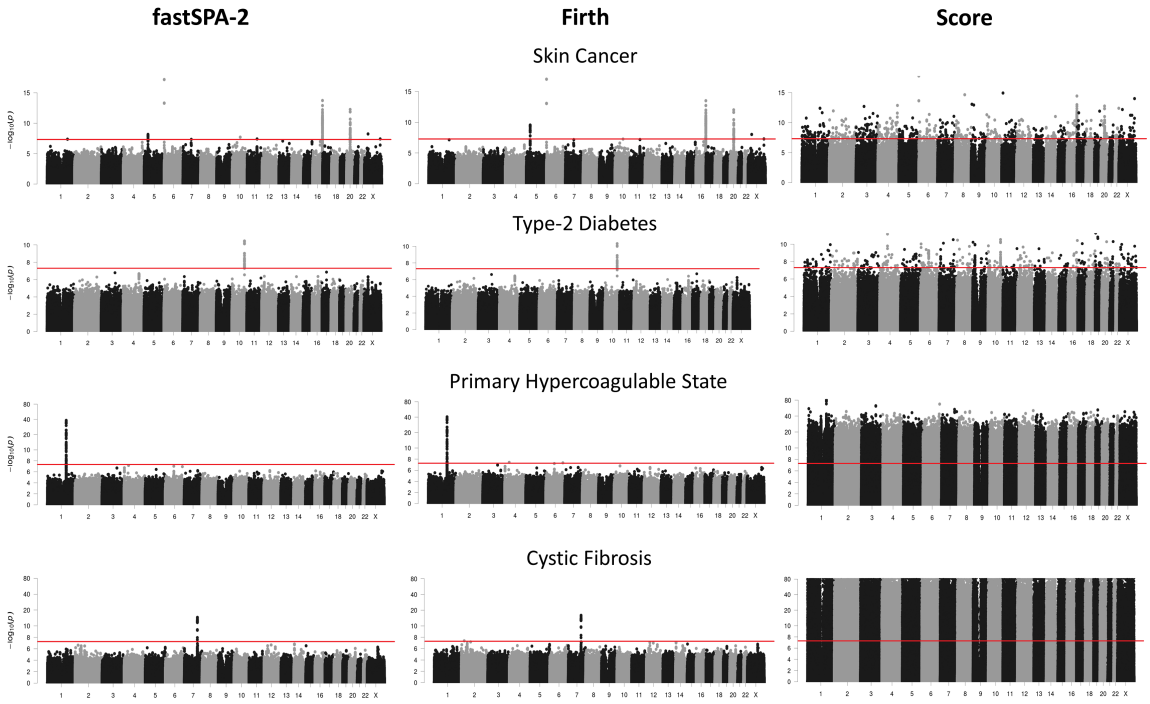


Figure E.3: Manhattan plots for four different phenotypes from the MGI data (all genotyped and imputed variants with minor allele count > 3 included). From left to right, the three panels show associations based on the fastSPA-2, Firth, and traditional score tests. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

APPENDIX F

Finding Optimal Nodes for Hermite Splines

Within each study, we need to obtain the node points at which we will share the functional values of $K'^{(j)}$ and $K''^{(j)}$, the first and second derivatives of the CGF of the score statistic in the j^{th} study under the null hypothesis. These node points can be calculated using any standard optimization algorithm (Nelder-mead (*Nelder and Mead*, 1965), variable metric (*Nocedal and Wright*, 2006), conjugate gradient (*Fletcher and Reeves*, 1964), coordinate descent (*Wright*, 2015) etc.). As Hermite splines provide local control of the spline approximation, i.e, perturbation of one node does not change the whole interpolating curve, it only changes the cubic pieces adjacent to that node, the coordinate descent algorithm can be applied for finding the optimal nodes. In the R package we implemented a coordinate descent algorithm tailored to our specific problem to find the optimal nodes. The software also allows the nodes to vary between different studies. By default, we use seven nodes in our software, one of which is kept fixed at zero since the values $K^{(j)}(0) = K'^{(j)}(0) = 0$ are known, and we share the value of $K''^{(j)}(0)$ which is the variance of score statistic within the j^{th} study. To find the remaining six nodes, we use the following loss function,

$$L(t, \underset{\sim}{u}) = w(t) \left| K'^{(j)}(t) - \hat{K}'^{(j)}_{\underset{\sim}{u}}(t) \right|,$$

where $\hat{K}'^{(j)}_{\underset{\sim}{u}}(t)$ is the spline approximation of $K'^{(j)}$ given the set of nodes $\underset{\sim}{u}$. The weights $w(t)$ are selected to be diminishing with larger absolute values of t as those values correspond to almost flatter or linear parts of the $K'^{(j)}$ function (see Figures F.1, F.2 and F.3). In simulations and real data analyses, we use the weights $w(t) = \min\left\{1, |t|^{-1/3}\right\}$, and approximate the total loss $R(\underset{\sim}{u})$ by summing the loss function $L(t, \underset{\sim}{u})$ over a grid of values of t. Then, to obtain the optimal set of nodes, we minimize the total loss using a coordinate descent algorithm where the nodes are treated as coordinates with the coordinate corresponding to the node at zero being fixed.

The calculation of total loss $R(\underset{\sim}{u})$ requires to evaluate the $K'^{(j)}$ function over a grid of t values, which involves $O(n_j)$ (or $O(m_j)$ when using the faster approximation) computations, where n_j is the sample size, and m_j is the number of minor allele carriers for the j^{th} study. Therefore, calculating optimal nodes for all of the variants can be computationally expensive. Since $\mu_i^{(j)}$s remain the same across all variants for the j^{th} study, the $K'^{(j)}$ function depends mainly on the MAF of the variant. To reduce the computation time, we group variants by their MAFs and obtain optimal nodes for each group. We first divide the MAF range $(0, 0.5]$ into 100 equal length bins, and randomly sample 100 variants within each bin. Then we compute the optimal nodes for those 100 selected variants, and use their coordinate-wise average of the nodes as fixed nodes for all the variants in that bin. In our real data application, we compared the p values when optimal nodes were calculated for all the variants, with p values when optimal nodes were calculated based only on 100 variants for each MAF bin to show that this reduced computation approach can still provide accurate p values. Figure F.4 shows that the p values from the two node-finding strategies were almost identical.

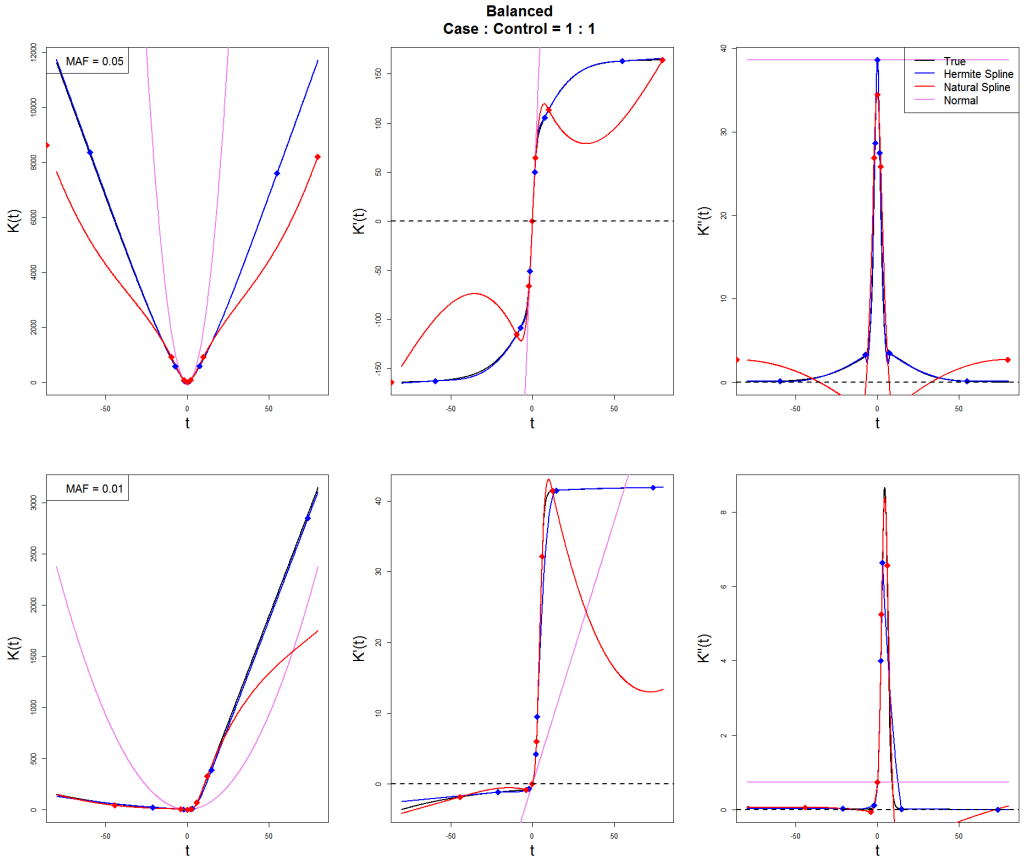


Figure F.1: Example of different spline and normal approximation curves in approximating the CGF and its derivatives for a study with 2000 samples and a balanced case-control ratio (1 : 1). The top and bottom panels represent variants with MAF = 0.05 and 0.01, respectively. From left to right, the plots show the curves of the CGF (K), its first derivative (K'), and its second derivative (K''). For the Hermite and natural spline approximations, seven optimal nodes (including a node at zero) were selected using the coordinate-descent algorithm described in Chapter V. The optimal nodes are represented by the diamond-shaped dots.

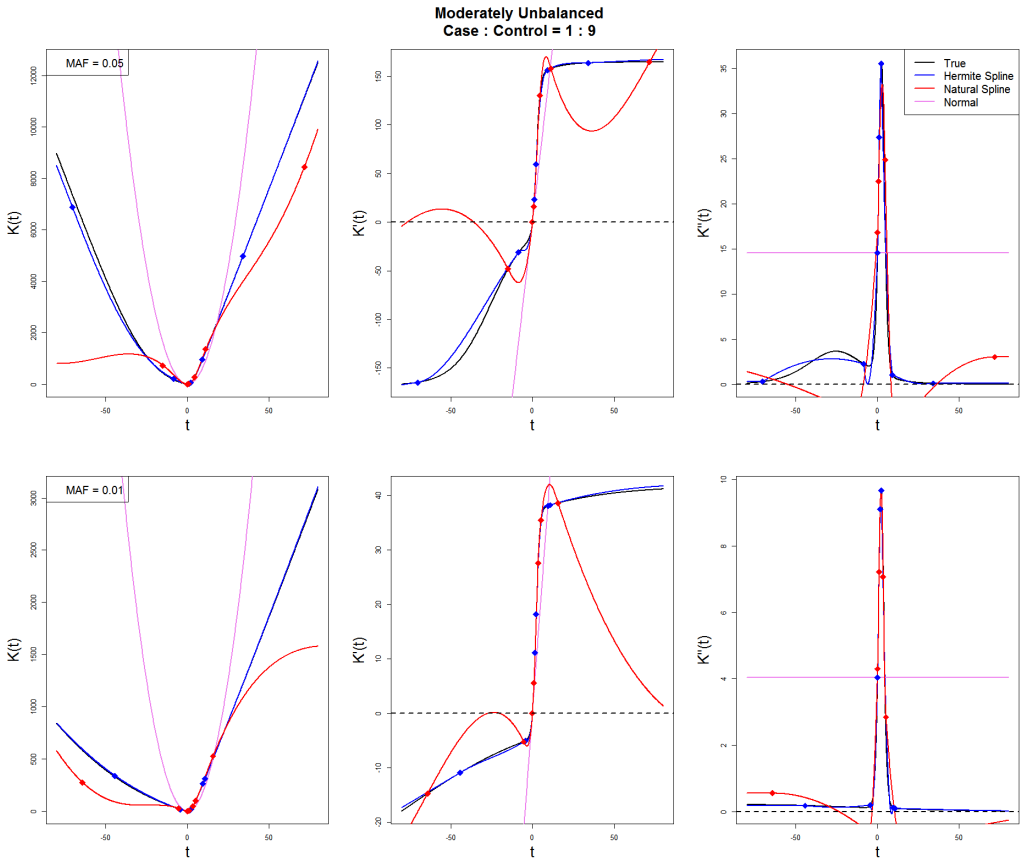


Figure F.2: Example of different spline and normal approximation curves in approximating the CGF and its derivatives for a study with 2000 samples and a moderately unbalanced case-control ratio ($1:9$). The top and bottom panels represent variants with MAF $= 0.05$ and 0.01, respectively. From left to right, the plots show the curves of the CGF (K), its first derivative (K'), and its second derivative (K''). For the Hermite and natural spline approximations, seven optimal nodes (including a node at zero) were selected using the coordinate-descent algorithm described in Chapter V. The optimal nodes are represented by the diamond-shaped dots.

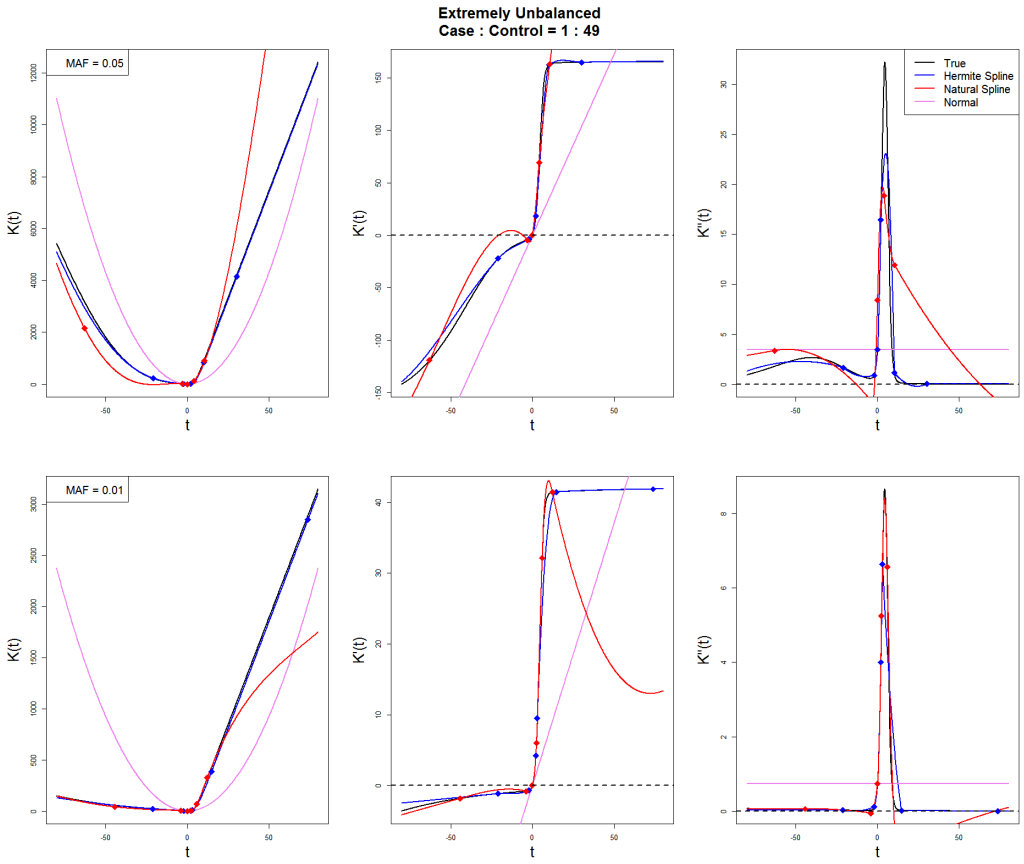


Figure F.3: Example of different spline and normal approximation curves in approximating the CGF and its derivatives for a study with 2000 samples and a extremely unbalanced case-control ratio (1 : 49). The top and bottom panels represent variants with MAF $= 0.05$ and 0.01, respectively. From left to right, the plots show the curves of the CGF (K), its first derivative (K'), and its second derivative (K''). For the Hermite and natural spline approximations, seven optimal nodes (including a node at zero) were selected using the coordinate-descent algorithm described in Chapter V. The optimal nodes are represented by the diamond-shaped dots.

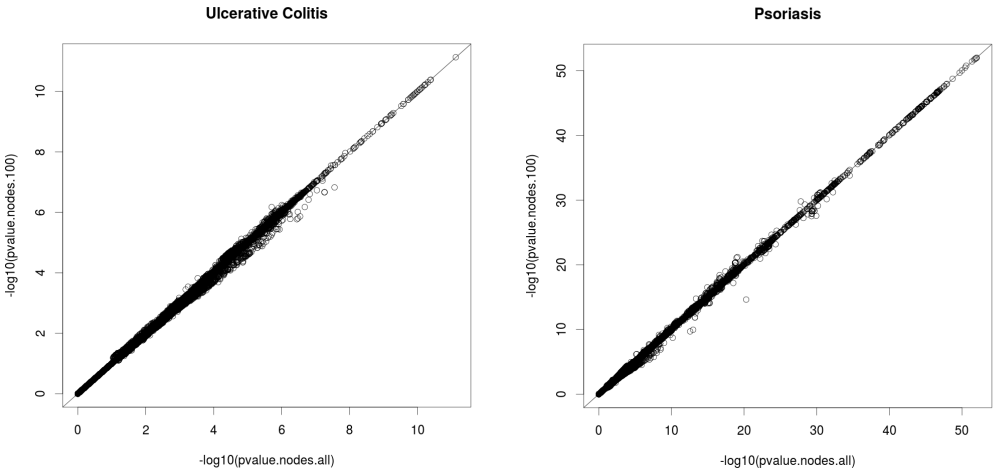


Figure F.4: Comparison of p values from the CGF-Spline method when using the node-finding algorithm for all variants against the reduced computation approach using the node finding algorithm only for 100 variants per MAF group. The left and right plots consider the phenotypes Ulcerative Colitis and Psoriasis, respectively from the UK Biobank interim release data. The X-axis represents the -log10 p values for the CGF-Spline method when optimal nodes were calculated for all of the variants, and the Y-axis represents the $-\log_{10}$ p values for the CGF-Spline method when optimal nodes were only calculated for 100 variants for each of the 100 MAF groups.

APPENDIX G

Simulation Details for Simulation Study 2 (Trans-Ethnic Meta-Analysis) in Chapter V

First, we selected 15000 samples of White ancestry (from a total of $\sim$ 140K unrelated samples) of the UK Biobank interim release data without replacement, and divided them in five groups of 3000 samples each. The samples with South Asian (2179 unrelated samples) and Black ancestry (1788 unrelated samples) in the data were placed into two groups of their own. Since this simulation is for imitating real data scenarios, and not real data analysis itself, we used the self-reported ancestries for this purpose. Within each of the seven groups, we performed principal component analysis to calculate the PC scores.

Phenotype Simulation

We simulated the phenotypes using the logistic regression model,

$$\text{logit}\left[\Pr\left(Y_i^{(j)}=1\right)\right]=\alpha^{(j)}+0.5\times\sum_{k=1}^{5}X_{ki}^{(j)}+G_i^{(j)}\gamma^{(j)}, \tag{G.1}$$

for $i=1,\ldots,n_j$, where $G_i^{(j)}$ is the genotype, $X_{ki}^{(j)}$s for $k=1,\ldots,4$ are the four PC scores for the i^{th} subject in the j^{th} group, and $X_{5i}^{(j)}\sim$ Bernoulli(0.5). To simulate the

phenotypes under the null hypothesis of no association, $\gamma^{(j)}$ was set to be zero. To simulate the phenotypes from different case-control imbalance settings, we selected $\alpha^{(j)}$ s corresponding to the prevalence of 0.4, 0.1 and 0.05, respectively for the case-control ratios of 1 : 1, 1 : 9 and 1 : 49. For the five groups with White ancestry samples, and the group with South Asian ancestry samples, we selected 2000 samples from each group, and from the Black ancestry group we selected 1500 samples, to match the case-control ratio. We denoted these seven sets of selected samples as the seven studies as mentioned in Chapter V.

Genotype Simulation

Within each study, the genotypes were simulated using $G_i^{(j)} \sim \text{Binomial}\left(2, \hat{\mu}_{\text{i}}^{(\text{j})}\right)$, where $\hat{\mu}_i^{(j)}$ is the fitted probability of success for the i^{th} subject in the j^{th} study under the binomial regression model,

$$\eta_i^{(j)} = \alpha_1^{(j)} + \sum_{k=1}^{4} X_{ki}^{(j)} \beta_{1k}^{(j)}.$$

$\eta_i^{(j)} = \exp\left(\mu_i^{(j)}\right) / \left[1 + \exp\left(\mu_i^{(j)}\right)\right]$, where the genotypes $G_i^{(j)} \sim \text{Binomial}\left(2, \mu_{\text{i}}^{(\text{j})}\right)$. To estimate the coefficients $\alpha_1^{(j)}$ and $\beta_{1k}^{(j)}$ for $k = 1, \ldots, 4$, we regressed the genotypes on the PC scores within each group for variants which have MAF at least 0.001 in all groups. Instead of directly using binomial regression, we first applied linear regression and then converted the linear regression coefficients to binomial regression coefficients, because linear regression is computationally much faster than binomial regression. To convert the linear regression coefficients to binomial regression coefficients, we used a method similar to the one described by *Pirinen et al.* (2013). In that paper, the authors described the method to estimate the logistic regression coefficients (with Bernoulli outcomes) using linear regression coefficients. In our case, the genotypes are assumed to follow binomial distributions instead of Bernoulli distributions typically

assumed in logistic regression models. However, since the canonical link function for both Bernoulli and binomial outcomes are the same (logistic link function), a simple modification of this method will serve our purpose. To apply our modified method, we first express the linear regression model as

$$G_i^{(j)} = \alpha_0^{(j)} + X_i^{(j)}\beta_0^{(j)} + \epsilon_i^{(j)},$$

and the binomial regression model as

$$\eta_i^{(j)} = \alpha_1^{(j)} + X_i^{(j)}\beta_1^{(j)}, \tag{G.2}$$

where $X_i^{(j)} = \left(X_{1i}^{(j)}, \ldots, X_{4i}^{(j)}\right)^T$, $\beta_0^{(j)} = \left(\beta_{01}^{(j)}, \ldots, \beta_{04}^{(j)}\right)$, $\beta_1^{(j)} = \left(\beta_{11}^{(j)}, \ldots, \beta_{14}^{(j)}\right)$ and $\epsilon_i^{(j)}$ is the error term for the linear regression model. We want to estimate the logistic regression coefficients $\alpha_1^{(j)}$ and $\beta_1^{(j)}$ using the linear regression coefficient estimates $\hat{\alpha}_0^{(j)}$ and $\hat{\beta}_0^{(j)}$. Then, under the binomial regression model (G.2) the Taylor series expansion of $E\left(G_i^{(j)}\right)$ around the mean is given by,

$$\begin{aligned} E\left(G_i^{(j)}\right) = 2\mu_i^{(j)} = 2\frac{e^{\alpha_1^{(j)} + X_i^{(j)}\beta_1^{(j)}}}{1+e^{\alpha_1^{(j)} + X_i^{(j)}\beta_1^{(j)}}} &= 2\frac{e^{\alpha_1^{(j)}}}{1+e^{\alpha_1^{(j)}}} + 2\frac{e^{\alpha_1^{(j)}}}{\left[1+e^{\alpha_1^{(j)}}\right]^2}X_i^{(j)}\beta_1^{(j)} \\ &+ \frac{e^{\alpha_1^{(j)}}\left(1-e^{\alpha_1^{(j)}}\right)}{\left[1+e^{\alpha_1^{(j)}}\right]^3}\left(X_i^{(j)}\beta_1^{(j)}\right)^2 + \ldots \end{aligned}$$

Assuming the effects in log-odds scale to be small such that $\left(X_i^{(j)}\beta_1^{(j)}\right)^2 \approx 0$, the first order approximation of $E\left(G_i^{(j)}\right)$ is given by,

$$E\left(G_i^{(j)}\right) \approx 2\frac{e^{\alpha_1^{(j)}}}{1+e^{\alpha_1^{(j)}}} + 2\frac{e^{\alpha_1^{(j)}}}{\left[1+e^{\alpha_1^{(j)}}\right]^2}X_i^{(j)}\beta_1^{(j)}. \tag{G.3}$$

This expression is of the same form as in the expression of $E\left(G_i^{(j)}\right)$ in the linear regression model,

$$E\left(G_i^{(j)}\right) = \alpha_0^{(j)} + X_i^{(j)}\beta_0^{(j)} \tag{G.4}$$

Comparing (G.3) and (G.4), we get the relationship between the logistic and linear regression coefficients given by,

$$\alpha_0^{(j)} = 2\frac{e^{\alpha_1^{(j)}}}{1+e^{\alpha_1^{(j)}}}, \qquad \beta_0^{(j)} = 2\frac{e^{\alpha_1^{(j)}}}{\left[1+e^{\alpha_1^{(j)}}\right]^2}\beta_1^{(j)}.$$

Therefore, the logistic regression coefficient estimates $\hat{\alpha}_1^{(j)}, \hat{\beta}_1^{(j)}$ can be calculated from the linear regression estimates $\hat{\alpha}_0^{(j)}, \hat{\beta}_0^{(j)}$ using the formulae,

$$\hat{\alpha}_1^{(j)} = \log\left(\frac{\hat{\alpha}_0^{(j)}/2}{1-\hat{\alpha}_0^{(j)}/2}\right), \qquad \hat{\beta}_1^{(j)} = \frac{\hat{\beta}_0^{(j)}/2}{\hat{\alpha}_1^{(j)}\left(1-\hat{\alpha}_1^{(j)}\right)}.$$

APPENDIX H

UK Biobank Data Description

Interim Release Data

The individuals were genotyped based on UK BiLEVE Axiom array (*Wain et al.*, 2015) (807,411 markers, $\sim$ 50K individuals genotyped) and UK Biobank Axiom array (*The UK Biobank Array Design Group*, 2014) (825,927 markers, $\sim$ 100K individuals genotyped) by Affymetrix, and then genotypes were imputed using a combined reference panel of UK10K (*The UK10K Consortium et al.*, 2015) and 1000 Genomes Phase 3 (*The 1000 Genomes Project Consortium et al.*, 2015) panels. The final genotype data contained 784,256 directly genotyped and $\sim$ 72 million imputed autosomal markers.

Data Pre-Processing

PheWAS codes (`https://phewascatalog.org/phecodes`) were used to denote the phenotypes. The White British samples were inferred using both self-reported information and genetic similarity, then the related samples were removed using KING

(*Manichaikul et al.*, 2010; *UK Biobank*, 2015) The principal components were calculated based on the samples within each study using the fastPCA (*Galinsky et al.*, 2016a,b; *Price et al.*, 2006) method. In the documentation of the UK Biobank interim release (*UK Biobank*, 2015), it was mentioned that 65 genotyped autosomal markers have significantly different allele frequencies in the UK BiLEVE and the UK Biobank Axiom arrays, 27 of which were included in phasing and imputation. We removed those 65 genotyped markers as well as any imputed markers within 10Kb neighbourhood of those 27 markers used for phasing and imputation, from our analysis.

APPENDIX I

Supplementary Tables and Figures for Chapter V

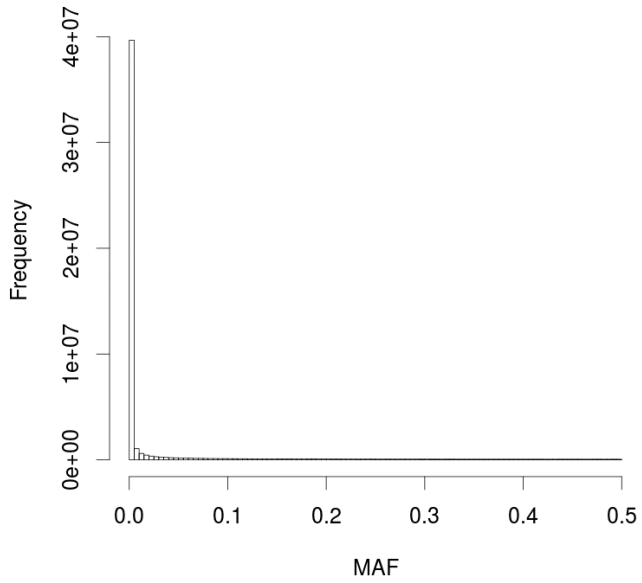


Figure I.1: Histogram of MAFs based on the white British ancestry samples from the UK Biobank interim release data.

Center	Ulcerative Colitis (555.2)			Psoriasis (696.4)		
	Case	Control	Total	Case	Control	Total
Manchester	32	2250	2282	23	2953	2976
Oxford	27	2589	2616	12	2992	3004
Cardiff	35	3660	3695	21	4116	4137
Glasgow	53	3463	3516	28	4496	4524
Edinburgh	28	3755	3783	15	4034	4049
Stoke	38	4187	4225	29	4738	4767
Reading	52	5815	5867	20	6666	6686
Bury	70	5891	5961	41	6947	6988
Newcastle	84	7234	7318	78	8581	8659
Leeds	96	8907	9003	72	10220	10292
Bristol	72	8883	8955	60	10047	10107
Barts	24	1630	1654	5	1905	1910
Nottingham	75	6759	6834	56	7759	7815
Sheffield	54	6225	6279	39	7229	7268
Liverpool	63	6468	6531	74	7464	7538
Middlesborough	49	4255	4304	20	4977	4997
Hounslow	32	4278	4310	18	4814	4832
Croydon	37	4020	4057	14	4666	4680
Birmingham	29	4480	4509	32	4939	4971
*Swansea	3	460	463	1	535	536
*Wrexham	2	169	171	1	191	192
*Stockport	0	64	64	0	82	82

Table I.1: Case-control sample sizes for Ulcerative Colitis and Psoriasis, across different assessment centers based on the unrelated samples from UK Biobank interim release data with white British ancestry. Swansea, Wrexham and Stockport were excluded from the analysis as those centers had less than five cases each, for each of the phenotypes.

Phenotype	Method	p value	Genomic control at the q^{th} p value quantile $q = 0.5$ (Median)	$q = 0.001$	$q = 0.001$
Ulcerative Colitis	Z-Score	Score	0.82	1.33	2.03
		fastSPA - 0.1	6.63	3.18	2.18
		fastSPA - 2	0.68	0.82	0.92
	GC	fastSPA - 0.1	1.40	0.96	1.00
		fastSPA -2	0.85	1.01	1.03
	CGF-Spline	N/A	0.74	1.00	1.02
Psoriasis	Z-Score	Score	0.71	1.54	2.51
		fastSPA - 0.1	9.14	3.53	2.41
		fastSPA - 2	0.58	0.80	0.98
	GC	fastSPA - 0.1	1.92	0.93	1.05
		fastSPA -2	0.73	1.01	1.09
	CGF-Spline	N/A	0.64	1.04	1.10
Psoriasis (MHC region excluded)	Z-Score	Score	0.71	1.51	2.36
		fastSPA - 0.1	9.14	3.52	2.4
		fastSPA - 2	0.58	0.78	0.88
	GC	fastSPA - 0.1	1.91	0.91	0.97
		fastSPA -2	0.73	1.00	1.02
	CGF-Spline	N/A	0.64	1.02	1.03

Table I.2: Estimated inflation factor of the genomic controls at different p value quantiles for different meta-analysis methods applied on the phenotypes Ulcerative Colitis and Psoriasis, from the UK Biobank interim release data. For Psoriasis, inflation factors were also calculated excluding the MHC region.

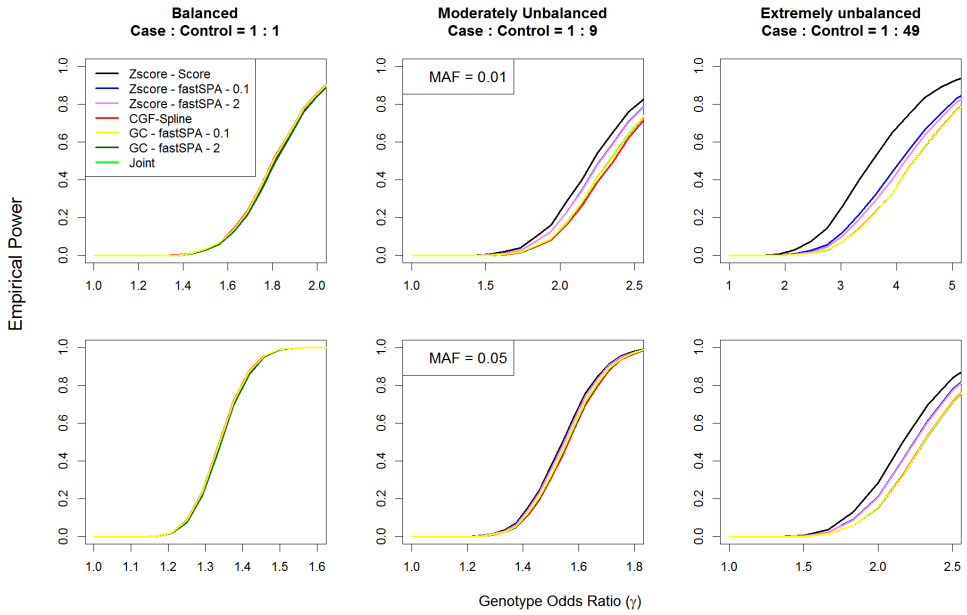


Figure I.2: Power curves for different meta-analysis methods at the nominal type I error level $\alpha = 5 \times 10^{-8}$ based on 5000 simulated datasets. Top panel considers MAF = 0.01 and bottom panel considers MAF = 0.05. From left to right, the plots consider case-control ratios 1 : 1, 1 : 9 and 1 : 49, respectively. In each plot the X-axis represents genotype odds ratios and the Y-axis represents the empirical power.

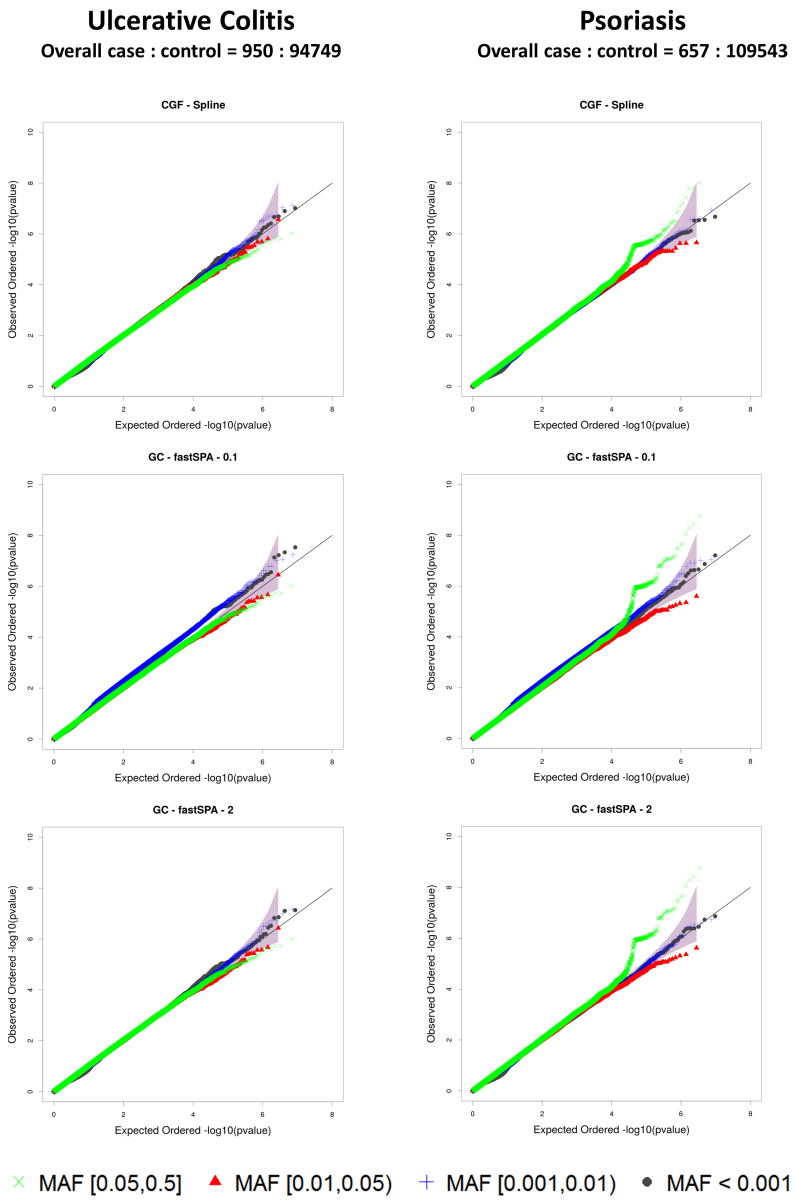


Figure I.3: QQ plots for our proposed methods when the within-study tests were performed on the imputed dosages from the UK Biobank interim release data. The left panel corresponds to Ulcerative Colitis, and the right panel corresponds to Psoriasis. Known associated loci in the MHC region were removed from the QQ plots. The plots are color-coded based on different MAF categories. To apply the GC method, genotype counts were calculated based on the best-called genotypes. We also calculated the genotype counts from MAFs using the Hardy-Weinberg equilibrium, and by rounding the dosage values to the nearest integer. In both situations, the QQ plots for the GC method were almost identical to the ones presented here, and hence were omitted.

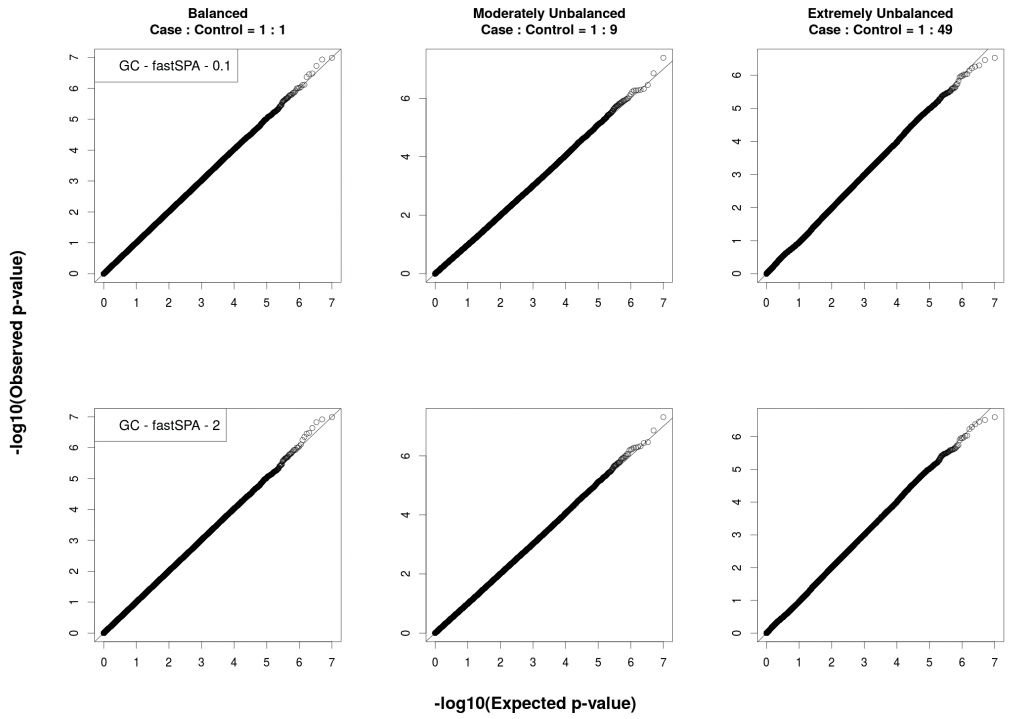


Figure I.4: QQ plots for the genotype count-based method using numerical simulations with very strong covariate effects. The simulation setting is identical to the simulation study 1 discussed in the main manuscript, except the log-odds ratios of the non-genetic covariates are set at 1.5 instead of 0.5. P values were obtained from 10 million simulated datasets with MAFs selected randomly from the MAF spectrum (Figure I.1) of the white British ancestry samples from the UK Biobank interim release data.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan (2010), Data quality control in genetic case-control association studies, *Nature protocols*, *5*(9), 1564–1573.

Anderson, C. A., et al. (2011), Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47, *Nature genetics*, *43*(3), 246–252, doi:10.1038/ng.764.

Bai, Z., and J. Yao (2012), On sample eigenvalues in a generalized spiked population model, *Journal of Multivariate Analysis*, *106*, 167 – 177, doi: http://dx.doi.org/10.1016/j.jmva.2011.10.009.

Bai, Z. D., and J. W. Silverstein (1998), No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices, *Ann. Probab.*, *26*(1), 316–345, doi:10.1214/aop/1022855421.

Baik, J., and J. W. Silverstein (2006), Eigenvalues of large sample covariance matrices of spiked population models, *Journal of Multivariate Analysis*, *97*(6), 1382 – 1408, doi:http://dx.doi.org/10.1016/j.jmva.2005.08.003.

Barndorff-Nielsen, O. E. (1990), Approximate interval probabilities, *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*(3), 485–496.

Barrett, J. H., et al. (2011), Genome-wide association study identifies three new melanoma susceptibility loci, *Nature genetics*, *43*(11), 1108—1113, doi: 10.1038/ng.959.

Bartels, R. H., J. C. Beatty, and B. A. Barsky (1987), *An introduction to splines for use in computer graphics and geometric modeling*, xiv, 476 p. pp., M. Kaufmann Publishers, Los Altos, Calif.

Bastien, P., V. E. Vinzi, and M. Tenenhaus (2005), Pls generalised linear regression, *Computational Statistics & Data Analysis*, *48*(1), 17–46, doi: https://doi.org/10.1016/j.csda.2004.02.005.

Berry, A. C. (1941), The accuracy of the gaussian approximation to the sum of independent variates, *Transactions of the American Mathematical Society*, *49*(1-3), 122–136, doi:Doi 10.2307/1990053.

Bertina, R. M., B. P. Koeleman, T. Koster, F. R. Rosendaal, R. J. Dirven, H. de Ronde, P. A. van der Velden, and P. H. Reitsma (1994), Mutation in blood coagulation factor v associated with resistance to activated protein c, *Nature*, *369*(6475), 64–7, doi:10.1038/369064a0.

Boulesteix, A.-L., and K. Strimmer (2007), Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, *Briefings in Bioinformatics*, *8*(1), 32, doi:10.1093/bib/bbl016.

Boyd, S., and L. Vandenberghe (2004), *Convex Optimization*, Cambridge University Press, Cambridge.

Braak, H., and E. Braak (1996), Evolution of the neuropathology of alzheimer's disease, *Acta Neurol Scand Suppl*, *165*, 3–12.

Brent, R. P. (1973), *Algorithms for Minimization without Derivatives.*, Prentice-Hall, Englewood Cliffs, NJ.

Brereton, R., and G. R. Lloyd (2014), Partial least squares discriminant analysis: taking the magic away, *Journal of Chemometrics*, *28*(4), 213–225, doi: doi:10.1002/cem.2609.

Bycroft, C., et al. (2017), Genome-wide genetic data on 500,000 uk biobank participants, *bioRxiv 166298 doi:10.1101/166298*.

Carroll, R. J., L. Bastarache, and J. C. Denny (2014), R phewas: data analysis and plotting tools for phenome-wide association studies in the r environment, *Bioinformatics*, *30*(16), 2375–6, doi:10.1093/bioinformatics/btu197.

Chun, H., and S. Keleş (2010), Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *72*(1), 3–25, doi:10.1111/j.1467-9868.2009.00723.x.

Chung, M. K., K. J. Worsley, B. M. Nacewicz, K. M. Dalton, and R. J. Davidson (2010), General multivariate linear modeling of surface shapes using surfstat, *Neuroimage*, *53*(2), 491–505, doi:10.1016/j.neuroimage.2010.06.032.

Clementi, S., G. Cruciani, M. Pastor, A. M. Davis, and D. R. Flower (1997), Robust multivariate statistics and the prediction of protein secondary structure content, *Protein Eng.*, *10*(7).

Collins, F. S., and H. Varmus (2015), A new initiative on precision medicine, *New England Journal of Medicine*, *372*(9), 793–795, doi:doi:10.1056/NEJMp1500523.

Cooper, H. M., L. V. Hedges, and J. C. Valentine (2009), *The handbook of research synthesis and meta-analysis*, 2nd ed., xvi, 615 p. pp., Russell Sage Foundation, New York.

Cox, D., and D. Hinkley (1974), *Theoretical Statistics*, Chapman and Hall, London.

Crane, P. K., et al. (2012), Development and assessment of a composite score for memory in the alzheimer's disease neuroimaging initiative (adni), *Brain Imaging Behav*, *6*(4), 502–16, doi:10.1007/s11682-012-9186-z.

Dale, A. M., B. Fischl, and M. I. Sereno (1999), Cortical surface-based analysis. i. segmentation and surface reconstruction, *Neuroimage*, *9*(2), 179–94, doi:10.1006/nimg.1998.0395.

Daniels, H. E. (1954), Saddlepoint approximations in statistics, *Annals of Mathematical Statistics*, *25*(4), 631–650, doi:DOI 10.1214/aoms/1177728652.

Das, S., et al. (2016), Next-generation genotype imputation service and methods, *Nat Genet*, *48*(10), 1284–1287.

de Jong, S. (1993), Simpls: An alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263, doi:https://doi.org/10.1016/0169-7439(93)85002-X.

Delaneau, O., J.-F. Zagury, and J. Marchini (2013), Improved whole-chromosome phasing for disease and population genetic studies, *Nat Meth*, *10*(1), 5–6.

Denny, J. C., et al. (2010), Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, *Bioinformatics*, *26*(9), 1205–10, doi:10.1093/bioinformatics/btq126.

Denny, J. C., et al. (2011), Variants near foxe1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies, *Am J Hum Genet*, *89*(4), 529–42, doi:10.1016/j.ajhg.2011.09.008.

Dey, R., E. M. Schmidt, G. R. Abecasis, and S. Lee (2017), A fast and accurate algorithm to test for binary phenotypes and its application to phewas, *Am J Hum Genet*, *101*(1), 37–49, doi:10.1016/j.ajhg.2017.05.014.

Ding, B., and R. Gentleman (2005), Classification using generalized partial least squares, *Journal of Computational and Graphical Statistics*, *14*(2), 280–298, doi:10.1198/106186005X47697.

Ding, X. (2015), Convergence of sample eigenvectors of spiked population model, *Communications in Statistics - Theory and Methods*, *44*(18), 3825–3840, doi:10.1080/03610926.2013.833240.

El Karoui, N. (2008), Spectrum estimation for large dimensional covariance matrices using random matrix theory, *Ann. Statist.*, *36*(6), 2757–2790, doi:10.1214/07-AOS581.

Epstein, M. P., A. S. Allen, and G. A. Satten (2007), A simple and improved correction for population stratification in case-control studies, *Am J Hum Genet*, *80*(5), 921–30, doi:10.1086/516842.

Esseen, C. G. (1942), On the liapounoff limit of error in the theory of probability., *Ark. Mat. Astr. och Fys.*, *28A*(9), 1–19.

Esseen, C. G. (1956), A moment inequality with an application to the central limit theorem., *Skand. Aktuarietidskr.*, *39*, 160–170.

Evangelou, E., and J. P. Ioannidis (2013), Meta-analysis methods for genome-wide association studies and beyond, *Nat Rev Genet*, *14*(6), 379–89, doi:10.1038/nrg3472.

Feller, W. (1945), The fundamental limit theorems in probability, *Bull. Amer. Math. Soc.*, pp. 800–832.

Firth, D. (1993), Bias reduction of maximum likelihood estimates, *Biometrika*, *80*(1), 27–38, doi:10.1093/biomet/80.1.27.

Fischl, B., M. I. Sereno, and A. M. Dale (1999), Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system, *Neuroimage*, *9*(2), 195–207, doi:10.1006/nimg.1998.0396.

Fletcher, R., and C. M. Reeves (1964), Function minimization by conjugate gradients, *Computer Journal*, *7*(2), doi:DOI 10.1093/comjnl/7.2.149.

Fox, N. C., E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor (1996), Presymptomatic hippocampal atrophy in alzheimer's disease. a longitudinal mri study, *Brain*, *119 (Pt 6)*, 2001–7.

Fritsche, L. G., et al. (2018), Association of polygenic risk scores for multiple cancers in a phenome-wide study: Results from the michigan genomics initiative, *The American Journal of Human Genetics*, *102*(6), 1048–1061, doi:10.1016/j.ajhg.2018.04.001.

Galinsky, K. J., G. Bhatia, P. R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and A. L. Price (2016a), Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia, *Am J Hum Genet*, *98*(3), 456–472, doi:10.1016/j.ajhg.2015.12.022.

Galinsky, K. J., P. R. Loh, S. Mallick, N. J. Patterson, and A. L. Price (2016b), Population structure of uk biobank and ancient eurasians reveals adaptation at genes influencing blood pressure, *Am J Hum Genet*, *99*(5), 1130–1139, doi:10.1016/j.ajhg.2016.09.014.

Geladi, P., and B. R. Kowalski (1986), Partial least-squares regression: a tutorial, *Analytica Chimica Acta*, *185*, 1–17, doi:https://doi.org/10.1016/0003-2670(86)80028-9.

Gibbons, L. E., et al. (2012), A composite score for executive functioning, validated in alzheimer's disease neuroimaging initiative (adni) participants with baseline mild cognitive impairment, *Brain Imaging Behav*, *6*(4), 517–27, doi:10.1007/s11682-012-9176-1.

Girko, V. L. (1996), Strong law for the eigenvalues and eigenvectors of empirical covariance matrices, *Random Operators and Stochastic Equations*, *4*, 176–204, doi:10.1515/rose.1996.4.2.179.

Gromski, P. S., H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre (2015), A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding, *Anal Chim Acta*, *879*, 10–23, doi:10.1016/j.aca.2015.02.012.

Han, B., and E. Eskin (2011), Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies, *Am J Hum Genet*, *88*(5), 586–98, doi:10.1016/j.ajhg.2011.04.014.

He, Z., M. Zhang, S. Lee, J. A. Smith, S. L. R. Kardia, A. V. Diez Roux, and B. Mukherjee (2017), Set-based tests for the gene–environment interaction in longitudinal studies, *Journal of the American Statistical Association*, *112*(519), 966–978, doi:10.1080/01621459.2016.1252266.

Hebbring, S. J. (2014), The challenges, advantages and future of phenome-wide association studies, *Immunology*, *141*(2), 157–65, doi:10.1111/imm.12195.

Hebbring, S. J., S. J. Schrodi, Z. Ye, Z. Zhou, D. Page, and M. H. Brilliant (2013), A phewas approach in studying hla-drb1*1501, *Genes Immun*, *14*(3), 187–91, doi:10.1038/gene.2013.2.

Helland, I. S. (1990), Partial least squares regression and statistical models, *Scandinavian Journal of Statistics*, *17*(2), 97–114.

Howbrook, D. N., A. M. van der Valk, M. C. O'Shaughnessy, D. K. Sarker, S. C. Baker, and A. W. Lloyd (2003), Developments in microarray technologies, *Drug Discov Today*, *8*(14).

Huang, X., W. Pan, S. Grindle, X. Han, Y. Chen, S. J. Park, L. W. Miller, and J. Hall (2005), A comparative study of discriminating human heart failure etiology using gene expression profiles, *BMC Bioinformatics*, *6*(1), 205, doi:10.1186/1471-2105-6-205.

Huey, E. D., E. N. Goveia, S. Paviol, M. Pardini, F. Krueger, G. Zamboni, M. C. Tierney, E. M. Wassermann, and J. Grafman (2009), Executive dysfunction in frontotemporal dementia and corticobasal syndrome, *Neurology*, *72*(5), 453–459, doi:10.1212/01.wnl.0000341781.39164.26.

Jack, J., C. R., et al. (2008), The alzheimer's disease neuroimaging initiative (adni): Mri methods, *J Magn Reson Imaging*, *27*(4), 685–91, doi:10.1002/jmri.21049.

Jack, J., C. R., et al. (2010), Update on the magnetic resonance imaging core of the alzheimer's disease neuroimaging initiative, *Alzheimers Dement*, *6*(3), 212–20, doi:10.1016/j.jalz.2010.03.004.

Jacobs, L. C., et al. (2015), A genome-wide association study identifies the skin color genes irf4, mc1r, asip, and bnc2 influencing facial pigmented spots, *The Journal of investigative dermatology*, *135*(7), 1735–1742, doi:10.1038/jid.2015.62.

Jagust, W., A. Gitcho, F. Sun, B. Kuczynski, D. Mungas, and M. Haan (2006), Brain imaging evidence of preclinical alzheimer's disease in normal aging, *Ann Neurol*, *59*(4), 673–81, doi:10.1002/ana.20799.

Jensen, J. L. (1995), *Saddlepoint Approximations*, Oxford University Press, Oxford.

Johnstone, I. M. (2001), On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.*, *29*(2), 295–327, doi:10.1214/aos/1009210544.

Johnstone, I. M., and A. Y. Lu (2009), On consistency and sparsity for principal components analysis in high dimensions, *Journal of the American Statistical Association*, *104*(486), 682–693, doi:10.1198/jasa.2009.0121, pMID: 20617121.

Jung, S., and J. S. Marron (2009), Pca consistency in high dimension, low sample size context, *Ann. Statist.*, *37*(6B), 4104–4130, doi:10.1214/09-AOS709.

Kerem, B., J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui (1989), Identification of the cystic fibrosis gene: genetic analysis, *Science*, *245*(4922), 1073–80.

Kim, S., et al. (2013), Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel, *PLoS One*, *8*(7), e70,269, doi:10.1371/journal.pone.0070269.

Kreyszig, E. (2006), *Advanced engineering mathematics*, 9th ed., John Wiley, Hoboken, NJ.

Krokstad, S., A. Langhammer, K. Hveem, T. L. Holmen, K. Midthjell, T. R. Stene, G. Bratberg, J. Heggland, and J. Holmen (2013), Cohort profile: the hunt study, norway, *Int J Epidemiol*, *42*(4), 968–77, doi:10.1093/ije/dys095.

Kuonen, D. (1999), Saddlepoint approximations for distributions of quadratic forms in normal variables, *Biometrika*, *86*(4), 929–935, doi:DOI 10.1093/biomet/86.4.929.

Ledoit, O., and S. Péché (2010), Eigenvectors of some large sample covariance matrix ensembles, *Probability Theory and Related Fields*, *151*(1), 233–264, doi:10.1007/s00440-010-0298-3.

Lee, S., P. F. Sullivan, F. Zou, and F. A. Wright (2008), Comment on a simple and improved correction for population stratification, *The American Journal of Human Genetics*, *82*(2), 524 – 526, doi:https://doi.org/10.1016/j.ajhg.2007.10.014.

Lee, S., F. Zou, and F. A. Wright (2010), Convergence and prediction of principal component scores in high-dimensional settings, *Ann. Statist.*, *38*(6), 3605–3629, doi:10.1214/10-AOS821.

Lee, S., G. R. Abecasis, M. Boehnke, and X. Lin (2014a), Rare-variant association analysis: Study designs and statistical tests, *American Journal of Human Genetics*, *95*(1), 5–23, doi:10.1016/j.ajhg.2014.06.009.

Lee, S., F. Zou, and F. A. Wright (2014b), Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data, *Biometrika*, doi:10.1093/biomet/ast064.

Lee, S., C. Fuchsberger, S. Kim, and L. Scott (2015), An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies, *Biostatistics*, *17*(1), 1–15, doi:10.1093/biostatistics/kxv033.

Liu, F., et al. (2015), Genetics of skin color variation in europeans: genome-wide association studies with functional follow-up, *Human genetics*, *134*(8), 823–835, doi:10.1007/s00439-015-1559-0.

Loh, P.-R., et al. (2016), Reference-based phasing using the haplotype reference consortium panel, *Nat Genet*, *48*(11), 1443–1448.

Ma, C., T. Blackwell, M. Boehnke, L. J. Scott, and T. D. i. the Go (2013), Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants, *Genetic Epidemiology*, *37*(6), 539–550, doi:10.1002/gepi.21742.

Man, M. Z. (2004), Evaluating methods for classifying expression data, *J Biopharm Stat*, *14*(4), doi:10.1081/BIP-200035491.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen (2010), Robust relationship inference in genome-wide association studies, *Bioinformatics*, *26*(22), 2867–73, doi:10.1093/bioinformatics/btq559.

Marchini, J., and B. Howie (2010), Genotype imputation for genome-wide association studies, *Nat Rev Genet*, *11*(7), 499–511, doi:10.1038/nrg2796.

Marčenko, V. A., and L. A. Pastur (1967), Distribution of eigenvalues for some sets of random matrices, *Mathematics of the USSR-Sbornik*, *1*(4), 457.

Marx, B. D. (1996), Iteratively reweighted partial least squares estimation for generalized linear regression, *Technometrics*, *38*(4), 374–381, doi:10.2307/1271308.

McCarthy, S., et al. (2016), A reference panel of 64,976 haplotypes for genotype imputation, *Nat Genet*, *48*(10), 1279–1283.

McIntosh, A. R., F. L. Bookstein, J. V. Haxby, and C. L. Grady (1996), Spatial pattern analysis of functional brain images using partial least squares, *Neuroimage*, *3*(3 Pt 1), 143–57, doi:10.1006/nimg.1996.0016.

Mestre, X. (2006), On the asymptotic behavior of quadratic forms of the resolvent of certain covariance-type matrices., *Tech. rep.*, CTTC/RC/2006-001, Centre Tecnologic de Telecomunicacions de Catalunya.

Mestre, X. (2008a), Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates, *IEEE Transactions on Information Theory*, *54*(11), 5113–5129, doi:10.1109/TIT.2008.929938.

Mestre, X. (2008b), On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices, *IEEE Transactions on Signal Processing*, *56*(11), 5353–5368, doi:10.1109/TSP.2008.929662.

Monchi, O., H. Benali, J. Doyon, and A. P. Strafella (2008), Recent advances in neuroimaging methods, *International Journal of Biomedical Imaging*, *2008*, 218,582, doi:10.1155/2008/218582.

Moon, S. W., I. D. Dinov, J. Kim, A. Zamanyan, S. Hobel, P. M. Thompson, and A. W. Toga (2015), Structural neuroimaging genetics interactions in alzheimer's disease, *Journal of Alzheimer's disease : JAD*, *48*(4), 1051–1063, doi:10.3233/JAD-150335.

Nan, H., et al. (2009), Genome-wide association study of tanning phenotype in a population of european ancestry, *The Journal of investigative dermatology*, *129*(9), 2250–2257, doi:10.1038/jid.2009.62.

Nelder, J. A., and R. Mead (1965), A simplex-method for function minimization, *Computer Journal*, *7*(4), 308–313, doi:DOI 10.1093/comjnl/7.4.308.

Nho, K., et al. (2012), Voxel and surface-based topography of memory and executive deficits in mild cognitive impairment and alzheimer's disease, *Brain Imaging Behav*, *6*(4), 551–67, doi:10.1007/s11682-012-9203-2.

Nho, K., et al. (2013), Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment, *Mol Psychiatry*, *18*(7), 781–7, doi:10.1038/mp.2013.24.

Nho, K., et al. (2015), Protective variant for hippocampal atrophy identified by whole exome sequencing, *Ann Neurol*, *77*(3), 547–52, doi:10.1002/ana.24349.

Nocedal, J., and S. J. Wright (2006), Numerical optimization, second edition, *Numerical Optimization, Second Edition*, pp. 1–664, doi:10.1007/978-0-387-40065-5.

Pa, J., K. L. Possin, S. M. Wilson, L. C. Quitania, J. H. Kramer, A. L. Boxer, M. W. Weiner, and J. K. Johnson (2010), Gray matter correlates of set-shifting among neurodegenerative disease, mild cognitive impairment, and healthy older adults, *Journal of the International Neuropsychological Society : JINS*, *16*(4), 640–650, doi:10.1017/S1355617710000408.

Paul, D. (2007), Asymptotics of sample eigenstruture for a large dimensional spiked covariance model, *Statist. Sinica*, *17*(4), 1617–1642.

Pendergrass, S. A., et al. (2013), Phenome-wide association study (phewas) for detection of pleiotropy within the population architecture using genomics and epidemiology (page) network, *PLoS Genet*, *9*(1), e1003,087, doi: 10.1371/journal.pgen.1003087.

Pirinen, M., P. Donnelly, and C. C. A. Spencer (2013), Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies, *Ann. Appl. Stat.*, *7*(1), 369–390, doi:10.1214/12-AOAS586.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1992), *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2 ed., Cambridge University Press, Cambridge, England.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006), Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, *38*(8), 904–909, doi: 10.1038/ng1847.

Ritchie, M. D., et al. (2013), Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk, *Circulation*, *127*(13), 1377–85, doi: 10.1161/CIRCULATIONAHA.112.000604.

Rubakhin, S. S., E. V. Romanova, P. Nemes, and J. V. Sweedler (2011), Profiling metabolites and peptides in single cells, *Nat. Methods*, *8*(4), doi: 10.1038/nmeth.1549.

Rusinek, H., S. De Santi, D. Frid, W. H. Tsui, C. Y. Tarshish, A. Convit, and M. J. de Leon (2003), Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal mr imaging study of normal aging, *Radiology*, *229*(3), 691–6, doi:10.1148/radiol.2293021299.

Saykin, A. J., et al. (2010), Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans, *Alzheimers Dement*, *6*(3), 265–73, doi:10.1016/j.jalz.2010.03.013.

Scott, L. J., et al. (2006), Association of transcription factor 7-like 2 (tcf7l2) variants with type 2 diabetes in a finnish sample, *Diabetes*, *55*(9), 2649–53, doi: 10.2337/db06-0341.

Shameer, K., et al. (2014), A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects, *Hum Genet*, *133*(1), 95–109, doi:10.1007/s00439-013-1355-7.

Shen, L., et al. (2014), Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers, *Brain Imaging Behav*, *8*(2), 183–207, doi: 10.1007/s11682-013-9262-z.

Shevtsova, I. G. (2010), An improvement of convergence rate estimates in the lyapunov theorem, *Doklady Mathematics*, *82*(3), 862–864, doi: 10.1134/s1064562410060062.

Silverstein, J., and S. Choi (1995), Analysis of the limiting spectral distribution of large dimensional random matrices, *Journal of Multivariate Analysis*, *54*(2), 295 – 309, doi:http://dx.doi.org/10.1006/jmva.1995.1058.

Solovieff, N., C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller (2013), Pleiotropy in complex traits: challenges and strategies, *Nat Rev Genet*, *14*(7), 483–95, doi:10.1038/nrg3461.

Stoica, P., and T. Söderström (1998), Partial least squares: A first-order analysis, *Scandinavian Journal of Statistics*, *25*(1), 17–24.

Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis (2005), Significance analysis of time course microarray experiments, *Proceedings of the National Academy of Sciences of the United States of America*, *102*(36), 12,837–12,842, doi: 10.1073/pnas.0504609102.

Stuart, P. E., et al. (2015), Genome-wide association analysis of psoriatic arthritis and cutaneous psoriasis reveals differences in their genetic architecture, *Am J Hum Genet*, *97*(6), 816–36, doi:10.1016/j.ajhg.2015.10.019.

Sulem, P., et al. (2007), Genetic determinants of hair, eye and skin pigmentation in europeans, *Nature genetics*, *39*(12), 1443–1452, doi:10.1038/ng.2007.13.

Sun, H., G. Y. Chen, and S. Q. Yao (2013), Recent advances in microarray technologies for proteomics, *Chemistry & Biology*, *20*(5), 685 – 699, doi: https://doi.org/10.1016/j.chembiol.2013.04.009.

The 1000 Genomes Project Consortium, et al. (2015), A global reference for human genetic variation, *Nature*, *526*(7571), 68–74, doi:10.1038/nature15393.

The UK Biobank Array Design Group (2014), Uk biobank axiom array content summary, *Available from: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf*.

The UK10K Consortium, et al. (2015), The uk10k project identifies rare variants in health and disease, *Nature*, *526*(7571), 82–90, doi:10.1038/nature14962.

Thomann, P. A., P. Toro, V. Dos Santos, M. Essig, and J. Schroder (2008), Clock drawing performance and brain morphology in mild cognitive impairment and alzheimer's disease, *Brain Cogn*, *67*(1), 88–93, doi:10.1016/j.bandc.2007.11.008.

UK Biobank (2015), Genotyping and quality control of uk biobank, a large-scale, extensively phenotyped prospective resource, *Available from: http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf*.

Verma, A., et al. (2018), Phewas and beyond: The landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from geisinger, *Am J Hum Genet*, *102*(4), 592–608, doi:10.1016/j.ajhg.2018.02.017.

Wain, L. V., et al. (2015), Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (uk bileve): a genetic association study in uk biobank, *Lancet Respir Med*, *3*(10), 769–81, doi:10.1016/S2213-2600(15)00283-0.

Welter, D., et al. (2014), The nhgri gwas catalog, a curated resource of snp-trait associations, *Nucleic Acids Res*, *42*(Database issue), D1001–6, doi:10.1093/nar/gkt1229.

Whittaker, E. T., and G. Robinson (1967), *The Newton-Raphson Method.*, pp. 84–87, 4 ed., Dover, New York.

Williams, N., and R. N. Henson (2018), Recent advances in functional neuroimaging analysis for cognitive neuroscience, *Brain and Neuroscience Advances*, *2*, 2398212817752,727, doi:10.1177/2398212817752727.

Wold, H. (1966), *Estimation of principal components and related models by iterative least squares.*, pp. 391–420, Academic Press, New York.

Wold, H. (1982), *Soft modelling, the basic design and some extensions.*, pp. 1–54, north Holland Publishing Company, Amsterdam.

Wold, S., A. Ruhe, H. Wold, and I. W. J. Dunn (1984), The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743, doi:10.1137/0905052.

Wold, S., M. Sjöström, and L. Eriksson (2001), Pls-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130, doi:https://doi.org/10.1016/S0169-7439(01)00155-1.

Worley, B., and R. Powers (2013), Multivariate analysis in metabolomics, *Curr. Metabolomics*, *1*(1), doi:10.2174/2213235X11301010092.

Wright, S. J. (2015), Coordinate descent algorithms, *Mathematical Programming*, *151*(1), 3–34, doi:10.1007/s10107-015-0892-3.

Zhang, M., et al. (2013), Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in european americans, *Human molecular genetics*, *22*(14), 2948–2959, doi:10.1093/hmg/ddt142.

Zhou, W., et al. (2018), Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies, *Nature Genetics (In Press)*.