Metabolomics in Obesity and Methods for Analyzing Multiple Phenotypes

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:16121149

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Metabolomics in Obesity and Methods for Analyzing Multiple Phenotypes

Zhonghua Liu

A Dissertation Submitted to the Faculty of
The Harvard T.H. Chan School of Public Health
in Partial Fulfillment of the Requirements
for the degree of Doctor of Science
in the Departments of Epidemiology and Biostatistics
Harvard University
Boston, Massachusetts

May 2015
Metabolomics in Obesity and Methods for Analyzing Multiple Phenotypes

Abstract

In this dissertation, we first assess the associations between circulating metabolites and body mass index in two U.S. prospective cohorts. Then, we propose several methods to analyze multiple phenotypes in genetic association studies based on summary statistics.

In the Chapter 1, we investigated the associations between circulating metabolites and repeatedly measured body mass index (BMI) among Caucasian men and women. We employed linear mixed models with a random intercept term to account for the within-subject correlation and adjust for potential baseline confounders including baseline obesity status, smoking, physical activity, alcohol intake, total caloric intake, age at blood draw and follow-up years. We found that the associations were not heterogeneous between Caucasian men and women, and therefore we used fixed effect meta-analysis to combine evidences of associations from these two cohorts. We found that valine, leucine and isoleucine were positively associated with BMI, while acetylglycine was negatively associated with BMI.

In Chapter 2, we present linear mixed model based score tests for the detection of pleiotropic genetic variants that are associated with multiple correlated traits based on summary statistics. Our tests are robust to effect heterogeneity and correlation structures among multiple traits. We conducted simulation studies to compare the proposed methods with existing methods. We also applied our methods to a global lipids GWAS summary statistics data set and identified hundreds of novel genetic variants.

In Chapter 3, we propose a geometric perspective on the powers of principal component association tests based on GWAS summary statistics. Utilizing eigen-analysis of
the correlation matrix and asymptotic power analysis, we investigate when PCA is powerful and when it is not to detect the genetic signals. We further apply our methods to a global lipids level genome-wide association study data set and identify hundreds of novel genetic variants that were missed by conventional single-trait analysis approaches. Our results can help guide researchers to choose powerful PCAT methods in multiple phenotype association studies and also better interpret the association results.
Contents

Title page ................................................................. i
Abstract ................................................................. ii
Table of Contents ....................................................... iv
List of Figures ........................................................... vii
List of Tables ............................................................ viii
Acknowledgments ......................................................... ix

1 Plasma Metabolites and Body Mass Index in U.S. Caucasian Men and Women 1
  1.1 Introduction ....................................................... 2
  1.2 Methods .......................................................... 2
    1.2.1 Study Design and Population .............................. 2
    1.2.2 Plasma Metabolites Profiling ............................. 3
    1.2.3 Assessment of BMI and Baseline Covariates ............ 3
    1.2.4 Statistical Analysis ..................................... 4
  1.3 Results .......................................................... 5
    1.3.1 Baseline Characteristics ................................. 5
    1.3.2 Metabolites Associated with Body Mass Index ........... 6
  1.4 Discussion ....................................................... 7

2 Multiple Phenotype Association Tests Based on Summary Statistics in Genome-wide Association Studies 9
  2.1 Introduction ....................................................... 10
  2.2 Methods .......................................................... 12
    2.2.1 Information in Summary Statistics ...................... 12
List of Figures

1.1 Significant metabolites associated with BMI . . . . . . . . . . . . . . . . . . . 7
2.1 Sizes and powers of MPATs for homogeneous effects when $K = 100$ . . . . . . 20
2.2 Sizes and powers of MPATs for heterogeneous effects when $K = 100$ . . . . 21
2.3 Venn diagram for novel SNPs detected by MPATs . . . . . . . . . . . . . . . . . . 24
3.1 Rejection boundaries of PCATs part a . . . . . . . . . . . . . . . . . . . . . . . . . 36
3.2 Rejection boundaries of PCATs part b . . . . . . . . . . . . . . . . . . . . . . . . . 37
3.3 Sizes and powers of PCATs for homogeneous effects when $K = 40$ . . . . . . 46
3.4 Sizes and powers of PCATs for heterogeneous effects when $K = 40$ . . . . . 48
3.5 Venn diagram for novel SNPs detected by PCATs . . . . . . . . . . . . . . . . . . 51
## List of Tables

1.1 Baseline characteristics of the study subjects. .............................................. 5  
1.3 Meta-analysis results using fixed effect meta-analysis with inverse variance weights. ............................................................................................................ 7  
2.1 A subset of the top associated SNPs with lipids levels based on the $p$-values of mixFisher method. CHR refers to the chromosome number. SUM refers to the SUM test targeting for homogeneous effects. VC refers to the variance component test targeting for heterogeneous effects. And mixFisher refers to the Fisher’s method of combining two independent $p$-values, which is expected to be more robust than both SUM and VC. HDL, LDL, TC and TG refer to high and low density lipoprotein cholesterol, total cholesterol and triglycerides respectively. ................................................................. 23  
3.1 A subset of novel SNPs identified by PCAT methods which can be detected by $PC_1$ but not the other PCs. ................................................................. 50  
3.2 A subset of novel SNPs identified by PCAT methods which can be detected by $PC_4$ but not the other PCs. ................................................................. 50
Acknowledgments

I owe my gratitude to my thesis advisor in the department of biostatistics, Professor Xihong Lin, for being an excellent mentor, encouraging and supporting. Thanks to Professor Liming Liang for his guidance as well.

I would like to thank my dissertation committee members, Lu Qi, Peter Kraft and Tyler VanderWeele for their helpful comments and insights, and for always being supportive about the research. I would also like to thank Xihong Lin’s group, and in particular Seunggeun Lee, Yen-Tsung Huang for helpful discussions.

Finally, I thank my family and friends for constant encouragement and support.
Plasma Metabolites and Body Mass Index in U.S. Caucasian Men and Women

Zhonghua Liu, Yan Zheng, Qibin Qi, Frank Hu, Peter Kraft, Lu Qi, Brian Wolpin and Liming Liang

Department of Epidemiology and Biostatistics
Harvard T.H. Chan School of Public Health
1.1 Introduction

Worldwide, the proportion of adults with a body-mass index (BMI) of $25 \text{ kg/m}^2$ or greater increased between 1980 and 2013 from 28.8% to 36.9% in men, and from 29.8% to 38% in women (Ogden et al., 2014). Obesity has become a major contributor to the global and national burden of various diseases (Misra and Khurana, 2008). Excessive adiposity is a risk factor for morbidity and mortality from cardiovascular diseases, type 2 diabetes mellitus (T2D) and some cancers (Danaei et al., 2009; Whitlock et al., 2009). Weight gain is primarily attributable to positive energy balance which occurs when an individual’s caloric intake exceeds one’s energy expenditure. It has been postulated that inherent metabolic factors might play a critical role in regulating energy homeostasis and one’s susceptibility to obesity (Hu, 2008).

Recently, the advancement of high throughput technologies such as targeted liquid chromatography-tandem mass spectrometry (LC-MS) metabolite profiling platform can allow simultaneous identification and quantification of a large number of circulating metabolites, i.e., metabolomics. Plasma metabolite profiles could help identify novel biomarkers for obesity development and progression. This would provide public health intervention pathways to reduce the incidence of obesity. However, the relationship between circulating metabolites and BMI remains unclear. In this study, we assessed the associations between 83 circulating metabolites, measured by a targeted LC-MS metabolomics platform and repeatedly measured BMI among Caucasian men and women in two U.S. cohorts, Health Professional Follow-Up Study (HPFS) and Nurses’ Health Study (NHS).

1.2 Methods

1.2.1 Study Design and Population

Our study population included participants from two large prospective cohort studies: Health Professional Follow-Up Study (HPFS) and Nurses’ Health Study (NHS). HPFS was initiated in 1986 when 51,529 U.S. men aged 40-75 years old working in health profes-
sions completed a mailed biennial questionnaire (Rimm et al., 1990). NHS was initiated in 1976 when 121,700 female nurses aged 30-55 years old completed a mailed biennial questionnaire (Colditz et al., 1986; Colditz, 1995). For the current study, we used 1990 as baseline for NHS and 1994 as baseline for HPFS, when the blood samples were collected. The current analysis included 135 U.S. Caucasian men and 216 U.S. Caucasian women who were originally chosen as the matched control samples for incident pancreatic cancer cases in a recent study (Mayers et al., 2014). The study subjects were free of cancer at the time of blood collection and type 2 diabetes (T2D) and their blood samples were profiled for metabolite concentrations. The study was approved by Human Research Committee at Brigham and Women’s Hospital (Boston, MA) and study participants provided informed consent.

1.2.2 Plasma Metabolites Profiling

Blood samples in EDTA tubes in HPFS were collected from 18,225 men in 1994, and in Heparin tubes in NHS were collected from 32,826 women in 1990. Blood collection procedures in HPFS and NHS have been described in more detail previously (Hankinson et al., 1995; Wei et al., 2005; Mayers et al., 2014). Plasma metabolites were measured as peak areas by a targeted LC-MS metabolomics platform using reference standards of metabolites to determine chromatographic retention times, MS monitoring transitions, de-clustering potentials and collision energies at Broad Institute of the Massachusetts Institute of Technology and Harvard University (Cambridge, MA) (Wang et al., 2011). 133 targeted metabolites were measured and 50 metabolites were excluded from final analysis, because 32 of them have poor stability with processing delay, 13 of them have mean coefficient of variation (CV) > 25% and 5 of them have missing data for more than 10% of participants (Townsend et al., 2013; Mayers et al., 2014). Finally, 83 of them were included in the subsequent analysis.

1.2.3 Assessment of BMI and Baseline Covariates

In both cohorts, height and body weight were assessed by questionnaire at baseline and every two years thereafter for 12 years (1990-2002 in NHS, 1994-2006 in HPFS). Baseline
obesity status is binary variable which takes value 1 if the baseline BMI > 30 kg/m² and 0 otherwise, and it was used as a baseline covariate. The main outcome was the repeatedly measured BMI between 1992-2002 in NHS and 1996-2006 in HPFS. In total, there were six BMI measurements for each study participant. Dietary and life-style factors were obtained and derived from the biennial questionnaires (Rimm et al., 1991; Colditz et al., 1997). In particular, physical activity information was obtained from questionnaire by asking subjects the amount of time they spent on average per week on various physical activities. And from this information, weekly energy expenditure in metabolic equivalent task-hours (MET-hours) was calculated (Ainsworth et al., 1993). The validity of physical activity, height and weight assessments were described previously (Pirie et al., 1981; Rimm et al., 1990; Wolf et al., 1994). Alcohol consumption and total caloric intake were assessed by validated food frequency questionnaires (FFQs) (Giovannucci et al., 1991). Smoking status was obtained from the questionnaires as well.

1.2.4 Statistical Analysis

All metabolite peak areas were naturally logarithmically transformed and then standardized to improve normality and interpretability. For repeated measurements of BMI, linear mixed models with a random intercept term were employed to assess the associations between plasma metabolites and BMI while accounting for the within-subject correlations among repeated measurements of BMI. To adjust for potential confounders, the following baseline covariates were included in the linear mixed models: baseline obesity status as a dummy variable (obese, not obese), smoking status (never, past and current smokers) as dummy variables, physical activity (MET-hours) as a continuous variable, alcohol intake (g/day) as a continuous variable, total calorie intake (kcal/day) as a continuous variable, age (years) at blood draw and the followup years since baseline. Stringent Bonferroni correction was employed and the significance threshold was set to be 0.0006 (0.05/83). The statistical analysis was performed in each cohort separately, and the effect heterogeneity between the two cohorts was assessed using Cochran’s Q statistic (Cochran, 1954). To assess the robustness of the association results, we further include incident diabetic cases from NHS (11 subjects) and HPFS (20 subjects) after blood collection into the analysis.
and then performed the same statistical analysis to the augmented study samples. All the analyses were performed using R software version 3.1.1, and the linear mixed models were fitted using R package *nlme*.

## 1.3 Results

### 1.3.1 Baseline Characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>HPFS</th>
<th>NHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Study Subjects</td>
<td>135</td>
<td>216</td>
</tr>
<tr>
<td>Body mass index, $kg/m^2$</td>
<td>25.8 (±3.2)</td>
<td>25.1 (±4.4)</td>
</tr>
<tr>
<td>Height, inches</td>
<td>70.1 (±2.9)</td>
<td>64.5 (±2.6)</td>
</tr>
<tr>
<td>Age at blood draw, years</td>
<td>65.4 (±8.0)</td>
<td>59.9 (±6.2)</td>
</tr>
<tr>
<td>Physical activity, MET-hr/wk</td>
<td>42.2 (36.6)</td>
<td>17.9 (19.3)</td>
</tr>
<tr>
<td>Alcohol, g/day</td>
<td>13.4 (17.4)</td>
<td>7.4 (12.5)</td>
</tr>
<tr>
<td>Total calorie intake, kcal/day</td>
<td>1976.5 (±602.9)</td>
<td>1840.8 (±495)</td>
</tr>
<tr>
<td>Smoking status, %</td>
<td>53(39.3)</td>
<td>87 (40.3)</td>
</tr>
<tr>
<td>Never</td>
<td>69(51.1)</td>
<td>93 (43.1)</td>
</tr>
<tr>
<td>Past</td>
<td>13(9.6)</td>
<td>36 (16.7)</td>
</tr>
<tr>
<td>Fasting samples ($\geq 8$ hours), %</td>
<td>75 (55.5)</td>
<td>164 (75.9)</td>
</tr>
</tbody>
</table>

Table 1.1: Continuous variables presented as mean (± standard deviation), categorical variables as number of subjects (%). Prevalent and incident diabetic cases were excluded. All the samples are from white-ethnic group. Abbreviations: HPFS, Health Professionals Follow-up Study; NHS, Nurses Health Study; The unit of physical activity is metabolic equivalent task-hours per week.

The baseline characteristics of the study subjects were summarized in Table 1.1. There were 135 men and 216 women of European descent included in the current analysis. The average baseline BMI in men (measured in 1994, $25.8 \pm 3.2$ $kg/m^2$) and women (measured in 1990, $25.1 \pm 4.4$ $kg/m^2$) were similar to each other. The average ages at baseline were 65 in men and 60 in women, so men were 5 years older than women on average at the time of blood collection. This was a noticeable characteristics indicating that the BMI of the current study samples might have little change over the next 12 years of follow up period. On average, women had 17.9 and men had 42.2 MET-hours of physical activity per week. Men consumed more alcohol than women per day in the study samples. Approximately
40% of the subjects were never-smokers in both men and women.

1.3.2 Metabolites Associated with Body Mass Index

We first analyzed the associations of metabolites with BMI in NHS and HPFS separately. We didn’t find any significant association heterogeneity between men and women based on Cochran’s Q testing statistics ($p > 0.05$). So we employed fixed effect meta-analysis with inverse-variance weights to combine the association results from NHS and HPFS. We summarized the cohort-specific and meta-analysis results of the significant metabolites associated with BMI ($p < 0.0006$) in Table 1.2 and Table 1.3 respectively. From the meta-analysis results, we found four significant metabolites associated with BMI. Three branched-chain amino acids (BCAA) were positively associated with BMI, and each one SD increases in the logarithms of valine, leucine and isoleucine were associated with $0.62 \pm 0.16 \text{ kg/m}^2$, $0.61 \pm 0.16 \text{ kg/m}^2$ and $0.57 \pm 0.16 \text{ kg/m}^2$ elevation in BMI. Amino acid acetylglycine was negatively associated with BMI and one SD increase in the logarithm of acetylglycine was associated with about $0.59 \pm 0.16 \text{ kg/m}^2$ decrease in BMI. A graphical presentation is also given in Figure 1.1.

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Group</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>valine</td>
<td>BCAA</td>
<td>0.52</td>
<td>0.23</td>
<td>0.02</td>
<td>0.71</td>
<td>0.23</td>
<td>0.002</td>
</tr>
<tr>
<td>leucine</td>
<td>BCAA</td>
<td>0.51</td>
<td>0.22</td>
<td>0.02</td>
<td>0.71</td>
<td>0.23</td>
<td>0.002</td>
</tr>
<tr>
<td>acetylglycine</td>
<td>Amino acids</td>
<td>-0.63</td>
<td>0.22</td>
<td>0.005</td>
<td>-0.55</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>isoleucine</td>
<td>BCAA</td>
<td>0.43</td>
<td>0.23</td>
<td>0.05</td>
<td>0.71</td>
<td>0.23</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 1.2: Associations between plasma metabolites and repeated measurements of BMI (1992-2002 in NHS and 1996-2006 in HPFS). Linear mixed models were used to assess the associations between plasma metabolites and repeated measurements of BMI, adjusted for baseline covariates including smoking status (never, past, current, missing), physical activity (MET-hour/week, continuous), alcohol intake (g/day, continuous), total calorie intake (kcal/day, continuous), age at blood draw (years, continuous), follow-up years and baseline obesity status. BCAA: branched-chain amino acid.

After including 11 and 20 incident diabetic cases from NHS and HPFS respectively into the statistical analysis, we found that the association results were very similar to the previous analysis results based on the samples without those incidence diabetic cases.
### Table 1.3: Meta-analysis results using fixed effect meta-analysis with inverse variance weights.

<table>
<thead>
<tr>
<th>Metabolites</th>
<th>Group</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
<th>Cochran’s p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>valine</td>
<td>BCAA</td>
<td>0.62</td>
<td>0.16</td>
<td>0.00015</td>
<td>0.56</td>
</tr>
<tr>
<td>leucine</td>
<td>BCAA</td>
<td>0.61</td>
<td>0.16</td>
<td>0.00017</td>
<td>0.53</td>
</tr>
<tr>
<td>acetylglycine</td>
<td>Amino acids</td>
<td>-0.59</td>
<td>0.16</td>
<td>0.00023</td>
<td>0.80</td>
</tr>
<tr>
<td>isoleucine</td>
<td>BCAA</td>
<td>0.57</td>
<td>0.16</td>
<td>0.00042</td>
<td>0.39</td>
</tr>
</tbody>
</table>

![Figure 1.1: Significant metabolites associated with BMI in U.S. Caucasian men and women.](image)

1.4 Discussion

In the current study, we found three branched-chain amino acids were positively associated with BMI and one amino acid acetylglycine was negatively associated with BMI in U.S. Caucasian men and women. The associations between those significant metabolites and BMI were not found to be heterogeneous between Caucasian men and women.

A number of prior studies have reported the positive associations between branched-chain amino acids (BCAA) valine, leucine and isoleucine (Newgard et al., 2009; Batch et al., 2013; Lackey et al., 2013) and BMI. The increased levels of BCAA in obese subjects were also reported previously (Adibi, 1968; Felig et al., 1969; Lackey et al., 2013). In particular, a recent study found that leucine-mediated cell signaling could help explain
the pathogenesis of obesity (Melnik, 2012). Other studies have also found that plasma BCAA concentrations typically increase in the context of obesity (Adibi, 1968; Felig et al., 1969; Lackey et al., 2013). Amino acid acetylglycine was found to be negatively associated with BMI for the first time. Although no direct previous evidence in the literature, however, acetylglycine was found to be associated with reduced risk of developing type 2 diabetes (Fiehn et al., 2010; Floegel et al., 2013). It could be postulated that the protective effects of increased levels of acetylglycine on the risk of diabetes might be through reducing BMI. Further studies are warranted to confirm this plausible pathway.

The major strength of our study is the prospective design with biennially measured BMI over 12 years period of follow-up since baseline in men and women. Since the first BMI used as outcome was two years after blood draw, therefore reverse causation bias is a minor issue in our study. Furthermore, we adjusted for important baseline covariates to mitigate confounding bias. To control for family-wise type I error rates, we used stringent Bonferroni corrected threshold to declare statistical significance. There might be more extra metabolites associated with BMI if we chose more liberal significance threshold, such as using false discovery rate (Benjamini and Hochberg, 1995). Several limitations of the present study need to be acknowledged. First, we used a targeted metabolic profiling platform, so there might be more metabolites measured by an untargeted platform that could be involved in the biological process of obesity. Second, there might be residual confounding factors need to be considered, for example, genetic and life style factors not measured in this study. Third, the sample size was not large so that we might not have enough powers to discover plasma metabolites related to BMI.

In conclusion, the present metabolic profiling study identified several circulating metabolites associated with BMI in Caucasian men and women in two independent prospective U.S. cohorts. Our findings provide potential biological pathways between circulating metabolites and BMI.
Multiple Phenotype Association Tests Based on Summary Statistics in Genome-wide Association Studies

Zhonghua Liu and Xihong Lin
Department of Epidemiology and Biostatistics
Harvard T.H. Chan School of Public Health
2.1 Introduction

To identify genetic variants underlying complex traits is the main goal of genome-wide association studies (GWASs), which have so far discovered thousands of genetic variants that are associated with hundreds of human traits and diseases (Hindorff et al., 2009). The rich results of GWAS studies have found that 4.6% of single nucleotide polymorphisms (SNPs) and 16.9% of genes are associated with multiple correlated phenotypes, indicating plausible biological pleiotropy (Solovieff et al., 2013). For example, a defect in the single gene that codes for phenylalanine hydroxylase results in multiple phenotypes associated with phenylketonuria (PKU) in humans, including mental retardation, eczema, and pigment defects (Paul, 2000). Purcell et al. (2009) also found that schizophrenia and bipolar disorder shared a substantial proportion of heritability. There remains a vast amount of genetic variants with pleiotropic effects to be discovered.

It is still a statistical challenge to identify such genetic variants and further to understand the underlying biological mechanism. Large sample size and more powerful statistical methods are both required to accomplish this task. Hence, more large multi-disease consortia like global lipids consortium need to be formed to pool data together to increase sample size, and more powerful and efficient statistical methods are demanded to analyze the pooled big data. An important issue might arise when analyzing pooled data is that individual level phenotype and genotype data could not be shared among consortia members and only summary statistics from single-trait analysis are available. Thus, the lack of individual level data renders several multiple phenotype regression-based methods, for examples, the ones proposed by (Roy et al., 2003; Klei et al., 2008; Ferreira and Purcell, 2009; Kim et al., 2009; O’Reilly et al., 2012; Schifano et al., 2013; Zhou and Stephens, 2014; Aschard et al., 2014) not applicable. We have to resort to summary statistics which actually contain rich information regarding the underlying genetic associations. How to best aggregate the information in the summary statistics to boost statistical powers to detect causal genetic variants remains unknown.

Note that the summary statistics for each genetic variant are correlated. Therefore, many traditional methods for combining independent testing statistics are not applicable, e.g.,
Fisher’s method for combining $p$-values (Fisher, 1932; Mosteller and Fisher, 1948), $Z$-score based methods (Stouffer et al., 1949), Pearson’s method (Pearson, 1934; Owen, 2009) and tail strength (Taylor and Tibshirani, 2006). Morton (1975), Kost and McDermott (2002) extended Fisher statistics for correlated tests by using scaled chi-squared distribution to approximate the exact distribution of Fisher statistics up to the first two moments. However, this approximation is not accurate enough in genetic association studies. Extensions of tail strength to combine dependent tests have been proposed (Wang and Shete, 2008), but the null distributions of the proposed tests were incorrectly derived (Zang et al., 2009; Han et al., 2009; Wang and Shete, 2009). A $p$-value based method for combining correlated tests proposed by (Conneely and Boehnke, 2007, 2010) is powerful when there are a few association signals with large genetic effects, but has substantial power loss when there exists a larger proportion of weak association signals. The methods proposed by (Xu et al., 2003; Yang et al., 2010; Zhu et al., 2015) don’t have analytic formulas for computing $p$-values and therefore might not be ideal for large-scale genetic association studies.

The objective of this paper is to propose robust, powerful and computationally efficient statistical testing procedures to detect genetic variants with pleiotropic effects based on summary statistics. We also show that the correlation structure among summary statistics at a genetic variant is the same as the correlation structure of the original continuous or binary phenotypes under the null hypothesis. We take the correlation structures explicitly into account for all of our methods. In particular, for homogeneous effects, we introduce an optimal score test that only tests one parameter instead of multiple ones, which greatly reduces the degrees of freedom. For heterogeneous effects, we propose a variance component test derived from a linear mixed model. In practice, we usually have no prior information about the underlying genetic effects, so we propose more robust testing procedures that can accommodate both homogeneous effects and heterogeneous effects by combining two independent score statistics that can capture the presence of shared group effects and/or individual effects. Specifically, we use inverse variance, inverse standard deviation weighting methods and an adaptive procedure to choose the linear combination coefficient for the two score statistics. We also propose Fisher’s and Tippett’s methods to combine $p$-values from these two score testing statistics. All of our
methods have analytic formulas to compute \( p \)-values and therefore are computationally efficient. We also carry out extensive simulation studies to empirically evaluate the type I error rates and powers in a wide range of situations. An application to a global lipids GWAS data set illustrates the usefulness of our methods.

2.2 Methods

2.2.1 Information in Summary Statistics

It’s a prerequisite to understand what kind of information is contained in the summary statistics before we develop novel testing procedures. Those summary statistics are collected from single-trait genome-wide association studies where multiple phenotypes (either continuous or binary) are analyzed separately. Suppose there are \( K \) traits and \( M \) genetic markers, and the genetic effect size and standard error output matrix can be denoted as \((\hat{\beta})_{mk} \) and \((\hat{se})_{mk} \), \( 1 \leq m \leq M, \ 1 \leq k \leq K \) respectively. We denote the Wald-type \( Z \)-testing statistics matrix as

\[
\left( Z \triangleq \frac{\hat{\beta}}{\hat{se}} \right)_{mk}.
\]

Each row of this summary statistics matrix is a random vector of length \( K \) with asymptotic multivariate normal distribution \( N(\mu, \Sigma_Z) \). In what follows, we will show how the summary statistics are related to the individual level data. We will use quantitative traits as an example, but we can apply the same argument for the binary traits as well.

Consider that a causal genetic variant affects \( K \) multiple correlated continuous traits, and single-trait genome-wide association studies have been performed for multiple phenotypes separately using the following normal linear regression model. For simplicity, we drop the subscript indexing the phenotype. For subject \( i \), we have

\[
Y_i = \beta_0 + \beta_G G_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2),
\]

where \( G_i \) represents the genotype of a single nucleotide polymorphism (SNP) for the \( i \)th subject. We can use maximum likelihood estimation procedure to estimate the regression coefficients and then use Wald-type \( Z \)-testing statistic for \( H_0 : \beta_G = 0 \) versus \( H_a : \beta_G \neq 0 \).
The mean of $Z$ is given by

$$E(Z) = \sqrt{n} \beta G \frac{\sigma_G \sqrt{n - 1} \Gamma((n - 2)/2)}{\sqrt{2} \Gamma((n - 1)/2)} \approx \sqrt{n} \beta G \frac{\sigma_G}{\sigma},$$

(2.2)

where $\sigma_G = \sqrt{2 \text{MAF}(1 - \text{MAF})}$ under the Hardy-Weinberg equilibrium and MAF represents the minor allele frequency of the SNP. Denote the correlation structure of $Z$ by $\Sigma_Z$ and denote the correlation matrix of the original multiple phenotypes by $R_Y$. We now show that these two correlation matrices are equal under the null. For simplicity, we assume $\sigma^2$ is known and denote the design matrix by $X$ with its first column all of ones and its second column the SNP genotype. Then, we have

$$\hat{\beta} = AY, A = (X^T X)^{-1} X^T, \text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

The genetic effect estimate is $\hat{\beta}_G = a_2^T Y$ and its standard error is $\text{se}_G = \sigma \sqrt{(a_2^T a_2)}$, where $a_2$ denotes the second row of matrix $A$. Assume that $Y^i$ and $Y^j$ are two different continuous traits of length $n$, and the corresponding Wald-type testing statistics are $Z^i$ and $Z^j$. Then we have

$$\text{Cov}(Z^i, Z^j) = \text{Cov}(a_2^T Y^i, a_2^T Y^j) = \frac{a_2^T a_2 \sigma^{i,j}}{\sqrt{a_2^T a_2 \sqrt{a_2^T a_2 \sigma^i \sigma^j}}} = \frac{\sigma^{i,j}}{\sigma^i \sigma^j},$$

where $\sigma^{i,j}, \sigma^i$ and $\sigma^j$ represent the covariance between the $i$th and $j$th phenotypes, and their standard deviations respectively. Therefore the covariance matrix of $Z$ for any genetic variant does not depend on the SNP genotype. This important result provides a way of estimating $\Sigma_Z$ by the sample covariance matrix of $Z$ over the independent null SNPs. Similar estimating procedures for the correlation matrix of the testing statistics can be found in (Yang et al., 2010; Stephens, 2013; Zhu et al., 2015).

### 2.2.2 Multiple Phenotype Association Tests (MPAT)

In this section, we will propose several methods for combining correlated univariate testing statistics to increase statistical power. In the following, we assume $\Sigma_Z$ is known or can be estimated consistently by its sample counterpart $\hat{\Sigma}_Z$, and we will drop its subscript for notation simplicity.
Testing for Homogeneous Effects

Under the assumption that the standardized effects $\mu^T = (\mu_1, \ldots, \mu_K)$ are of the same magnitudes and directions, the distribution of $Z$ is reduced to $N(\mu^* J, \Sigma)$, where $J^T = (1, 1, \ldots, 1)$ and $\mu^*$ is a scalar denoting the shared common effect. Then testing whether $H_0 : \mu = 0$ is equivalent to testing $H_0 : \mu^* = 0$. We propose a score test statistic called SUM given by

$$\text{SUM} = \frac{J^T \Sigma^{-1} Z}{\sqrt{J^T \Sigma^{-1} J}},$$

(2.3)

which follows $N(0, 1)$ under the null. It’s essentially a weighted sum of the components of $Z$ with weights equal to $J^T \Sigma^{-1}$, so a phenotype that is less correlated with the others have relatively higher weight. Moreover, it can be shown that SUM test is the most powerful test among all linear combinations of $Z$ in homogeneous effect settings (O’Brien, 1984; Wei and Johnson, 1985; Pocock et al., 1987; Xu et al., 2003; Bittman et al., 2009). An ad hoc method is to use the direct sum (DSUM) test statistic $\sum_{k=1}^K Z_k$ even when the phenotypes are correlated. Its null distribution is given by $N(0, J^T \Sigma J)$, so the $p$-value of DSUM can be computed easily.

Testing for Heterogeneous Effects

The effects of a genetic variant on multiple phenotypes might be of different magnitudes and directions. For example, a causal SNP could increase low density lipoprotein (LDL) cholesterol but decrease high density lipoprotein (HDL) cholesterol. Hence, the SUM test is expected to perform poorly in these situations. This motivates us to develop new testing procedures that can handle the situations in which the effects heterogeneous. We formulate this problem using the following model

$$Z = \mu + \epsilon, \ \epsilon \sim N(0, \Sigma), \ \mu^T = (\mu_1, \ldots, \mu_K).$$

(2.4)

We further assume that $\mu_k$ follow an arbitrary common distribution $F$ with mean 0 and variance $\tau$. Then testing whether $H_0 : \mu_1 = \cdots = \mu_K = 0$ is equivalent to testing whether $H_0 : \tau = 0$. Under this assumption, equation (2.4) becomes a linear mixed model (Laird and Ware, 1982). Following (Lin, 1997), one can derive the variance component score test
statistic for $\tau$ given by
\[ VC = Z^T \Sigma^{-1} \Sigma^{-1} Z, \tag{2.5} \]
which is a quadratic form in $Z$. It can be easily shown that $VC$ follows a mixture of chi-square distribution $\sum_{k=1}^{K} \lambda_k \chi^2_{1,k}$, where the weights $\lambda_k$ are the eigenvalues of the matrix $\Sigma^{-1}$. And the $p$-value of $VC$ can be calculated either by the exact method (Davies, 1980) or the moment matching method (Liu et al., 2009).

**Testing for Mixed Effects**

In practice, the underlying genetic effect patterns are usually unknown. Therefore, either SUM or VC test could lose substantial power if their model assumptions are violated. Hence, we need to develop more robust procedures that will perform well for all genetic effect patterns. Note that the genetic effects can be decomposed into group and individual effects, where the former captures the shared common genetic basis and the later characterizes the heterogeneity across different phenotypes. This effect decomposition leads to the following linear mixed model
\[ Z = \mu J + b + \epsilon, \quad \epsilon \sim N(0, \Sigma) \tag{2.6} \]
where $\mu$ is a scalar denoting the shared group effects and $b = (b_1, \ldots, b_K)^T$ are the individual random effects assumed to follow $N(0, \tau I_{KK})$. If $\tau = 0$, then this model reduces to the problem of testing for homogeneous effect and hence the SUM test could be used; if $\mu = 0$, then this model reduces to the case of testing for purely heterogeneous effects and hence the VC test could be used. Therefore, model (2.6) is more general and includes homogeneous and heterogeneous effect models as special cases.

Under model (2.6), testing associations between a genetic variant and the $K$ traits is equivalent to testing whether $H_0 : \mu = 0, \tau = 0$ jointly. To avoid intensive computation in likelihood ratio test or Wald test, score test provides computational advantage since it only involves fitting the model under the null. We now derive score test based on the following marginal model given by $Z \sim N(\mu 1, V)$, where $V = \tau I_{KK} + \Sigma$. The score for $\mu$ under $H_0 : \mu = 0, \tau = 0$ is given by $U_{\mu_0} = J^T \Sigma^{-1} Z$. The score for $\tau$, however is derived
under the modified \( H_0 \) without restricting \( \mu^* = 0 \), mimicking the idea proposed by (Sun et al., 2013), and is given by

\[
U_\tau = \frac{1}{2}(Z - \hat{\mu})^T \Sigma^{-1} \Sigma^{-1} (Z - \hat{\mu}) - \frac{1}{2} \text{tr}(\Sigma^{-1}),
\]

where \( \hat{\mu} \) is the maximum likelihood estimator of \( \mu \) under \( \tau = 0 \), which is simply the sample mean of \( Z \), i.e. \( \hat{\mu} = \frac{1}{K} \sum_{k=1}^{K} Z_k \). Because the second term in (2.7) does not depend on data, so we could use the first term to construct the test statistic under \( \tau = 0 \), given by

\[
U_{\tau_0} = (Z - \hat{\mu})^T \Sigma^{-1} \Sigma^{-1} (Z - \hat{\mu}).
\]

It can be shown that \( U_{\mu_0} \) and \( U_{\tau_0} \) are statistically independent, and so are \( U_{\mu_0}^2 \) and \( U_{\tau_0} \) (see appendix for proof). The variance of \( U_{\mu_0}^2 \) is \( 2\text{tr}[\Lambda_{\mu_0} \Sigma \Lambda_{\mu_0} \Sigma] \), where \( \Lambda_{\mu_0} = \Sigma^{-1} JJ^T \Sigma^{-1} \), and the variance of \( U_{\tau_0} \) is \( 2\text{tr} [\Lambda_{\tau_0} \Sigma \Lambda_{\tau_0} \Sigma] \), where \( \Lambda_{\tau_0} = (I - H) \Sigma^{-1} \Sigma^{-1} (I - H) \) and \( H = J(\Sigma^{-1/2} \Lambda \Sigma^{-1/2} J)^{-1} J^T \).

We now propose to combine these two independent score statistics into one given by

\[
T_\phi = \phi U_{\mu_0}^2 + (1 - \phi) U_{\tau_0} = Z^T \Lambda_\phi Z,
\]

where \( \Lambda_\phi = \phi \Lambda_{\mu_0} + (1 - \phi) \Lambda_{\tau_0} \) and \( \phi \in [0, 1] \). It can be easily shown that \( T_\phi \) follows a mixture of chi-squared distribution \( \sum_j \lambda_j \chi^2_j \) under the null, where \( \lambda_j \) are the eigenvalues of the matrix \( \Sigma^{1/2} \Lambda_\phi \Sigma^{1/2} \) for any given \( \phi \). And the \( p \)-value can be computed either by the exact method (Davies, 1980) or moment matching method (Liu et al., 2009).

However, \( \phi \) is usually unknown in practice and therefore needs to be chosen. According to minimum variance criterion, we propose the inverse-variance weighting scheme

\[
\phi = \frac{\text{Var}(U_{\tau_0})}{\text{Var}(U_{\mu_0}^2) + \text{Var}(U_{\tau_0})},
\]

and the resulting \( T_\phi \) is refereed to as mixVarWeight. We can also use inverse standard deviation weighting such that the two weighted testing statistics have equal variances, and this testing procedure is refereed to as mixSDWeight.

Instead of combining two testing statistics, we could combine the two independent \( p \)-values for \( U_{\mu_0} \) and \( U_{\tau_0} \) using Fisher’s method (Fisher, 1932) or Tippett’s procedure (Tip-
The p-value combination approaches are readily easy to implement and more importantly can help us identify the sources of associations, either attributable to shared group effect or individual effect, or both. It is therefore more informative for us to understand and interpret the genetic association results observed in the data. The implementation is straightforward. Denote $P_\mu$ and $P_\tau$ as the p-values from $U_{\mu_0}$ and $U_{\tau_0}$, then Fisher’s p-value is given by $P_{Fisher} = P(\chi^2_4 > -2\log(P_\mu) - 2\log(P_\tau))$, and Tippett’s $p$-value is given by $P_{Tippett} = 1 - (1 - \min(P_\mu, P_\tau))^2$. These two methods are called mixFisher and mix-Tippett respectively. To calculate each p-values, we need to establish the results about the distributions of the two testing statistics under the null hypothesis. Note that $P_\mu$ could be calculated easily from the null distribution of $U_{\mu_0}$ that is $U_{\mu_0} \sim N(0, J^T\Sigma^{-1}J)$. And $P_\tau$ can be calculated from the chi-square mixture distribution $\sum_j \lambda_j \chi^2_{1j}$, where $\lambda_j$ are the eigenvalues of the matrix $\Sigma^{1/2}(I - H)\Sigma^{-1}\Sigma^{-1}(I - H)\Sigma^{1/2}$.

**Adaptive Optimal Test**

We could also choose $\phi$ in an adaptive fashion and propose to use the minimum p-value as a testing statistic given by

$$P_{Ada} = \inf_{0 \leq \phi \leq 1} p_\phi,$$

where $p_\phi$ is the p-value computed based on $T_\phi$ for any fixed value of $\phi \in [0, 1]$. In practice, $P_{Ada}$ could be obtained by grid searching over a range of possible candidates in $[0, 1]$, denoted as $0 = \phi_1 < \phi_2 < \ldots < \phi_B = 1$, where $B$ is the number of grid points in the interval $[0, 1]$. If the observed value of $P_{Ada}$ is denoted as $P_{Ada}^{obs}$ then the p-value based on $P_{Ada}$ can be computed as

$$p-value = 1 - P(P_{Ada} > P_{Ada}^{obs}) \quad = 1 - P(p_{\phi_1} > P_{Ada}^{obs}, \ldots, p_{\phi_B} > P_{Ada}^{obs}) \quad = 1 - P(T_{\phi_1} < q(\phi_1), \ldots, T_{\phi_B} < q(\phi_B)),$$

where $q(\phi_b) = F_{T_{\phi_b}}^{-1}(1 - P_{Ada}^{obs})$ and $F_{T_{\phi_b}}^{-1}$ denotes the inverse cumulative distribution function of $T_{\phi_b}, 1 \leq b \leq B$. Note that the B-dimensional integration in (2.10) involves the knowledge about the joint distribution of $(T_{\phi_1}, \ldots, T_{\phi_B})$ which is difficult to obtain. The
following result shows that $T_\phi$ has the same asymptotic distribution as the random variable in (2.11),

$$\tau(\phi)\eta_0 + (1 - \phi)\kappa,$$  \hfill (2.11)

where $\tau(\phi) = \phi J^T\Sigma^{-1}J$ and $\eta_0 = \frac{U_{\mu^\gamma}}{J^T\Sigma^{-1}J}$ which has an asymptotic chi-squared distribution with one degree of freedom. And $\kappa = \sum \lambda_j \eta_j$, where $\lambda_j$ are the non-zero eigenvalues of the matrix $\Sigma^{1/2}\Lambda_0\Sigma^{1/2} = \Sigma^{1/2}(I - H)\Sigma^{-1}\Sigma^{-1}(I - H)\Sigma^{1/2}$, and $\eta_j$ are mutually independent $\chi^2_1$ random variables. Using this representation, we could compute the $p$-value in (2.10) using the law of total expectation rather than the original intractable $B$-dimensional integration. Therefore, equation (2.10) can be rewritten as

$$p\text{-value} = 1 - E\left[P(\kappa < \min_b \frac{q(\phi_b) - \tau(\phi_b)\eta_0}{1 - \phi_b} | \eta_0)\right]$$

$$= 1 - \int_0^\infty F_\kappa(\delta(x)) f_{\eta_0}(x) dx,$$  \hfill (2.12)

where $F_\kappa(\cdot)$ denotes the cumulative distribution function of $\kappa$ which is a mixture of chi-squared distributions defined above, $f_{\eta_0}(\cdot)$ is the probability density function of a standard $\chi^2_1$ random variable $\eta_0$, and $\delta(x)$ is

$$\delta(x) = \min_{1 \leq b \leq B} \frac{q(\phi_b) - \tau(\phi_b)x}{1 - \phi_b}.$$  \hfill (2.13)

So the $p$-value in equation (2.12) could be easily computed using one dimensional numerical integration. We provide the following algorithm to implement the proposed adaptive optimal testing scheme.

Step 1: Provide candidates for $\phi_b$ by choosing $B$ grid points in $[0, 1]$, denoted as $0 = \phi_1 < \phi_2 < \cdots < \phi_B = 1$.

Step 2: For each $\phi_b, 1 \leq b \leq B$, compute $T_{\phi_b}$ using the formulas given by

$$T_\phi = Z^T \Lambda_\phi Z, \quad \Lambda_\phi = \phi \Sigma^{-1}JJ^T\Sigma^{-1} + (1 - \phi)(I - H)\Sigma^{-1}\Sigma^{-1}(I - H).$$

Step 3: Compute $p_{\phi_b}$ using either exact method (Davies, 1980) or approximation methods like moment matching (Liu et al., 2009).
Step 4: Compute the following statistics

\[
P_{Ada} = \min_{1 \leq b \leq B} p_{\phi_b}
\]
\[
q(\phi_b) = F_{T_{\phi_b}}^{-1}(1 - P_{Ada})
\]
\[
\tau(\phi_b) = \phi_bJ^T\Sigma^{-1}J
\]
\[
\delta(x) = \min_{1 \leq b \leq B} \frac{q(\phi_b) - \tau(\phi_b)x}{1 - \phi_b}
\]

Step 5: Perform numerical integration to obtain \(p\)-value given by

\[
p\text{-value} = 1 - \int_0^\infty F_{\kappa}(\delta(x)) f_{\eta_0}(x)dx.
\]

2.3 Simulation studies

To evaluate the performances of the proposed methods, we conduct extensive simulation studies to compare their type I error rates and empirical powers in a wide range of situations. In particular, we consider the following factors: signal sparsity, effect heterogeneity and the correlation structure. In addition to the methods presented above, we also include WI test defined as \(Z^T Z\) and the traditional Wald-type test \(Z^T \Sigma^{-1} Z\) for comparison purpose.

The total number of phenotypes was set to be \(K = 100\) and the numbers of causal phenotypes were set to be \(K_1 = (1, 7, 25, 100)\). As for homogeneous genetic effects, we set \((\mu_1, \ldots, \mu_{K_1}, 0, \ldots, 0) = (1, \ldots, 1, 0, \ldots, 0) \times c\) where \(c \in [0, 4]\), and for heterogeneous genetic effects, we set half of the \(K_1\) effects to be negative and the other half to be positive. In the case \(K_1\) is odd, we set one more positive component, e.g., for \(K_1 = 25\), we set 13 positives and 12 negatives. For correlation structures, we considered three parameters \(\rho_1, \rho_2\) and \(\rho_3\) which represent the correlation among causal phenotypes, inter-correlation between causal and non-causal phenotypes, and the correlation among non-causal phenotypes respectively. More specifically, we set \(\rho_2 = (0, 0.1, 0.5)\). And for \(\rho_2 = 0\), we set \(\rho_1 = (0, 0.2, 0.8)\) and \(\rho_3 = (0, 0.2, 0.8)\). Similarly, for \(\rho_2 = 0.1\), we set \(\rho_1 = (0.1, 0.2, 0.8)\) and \(\rho_3 = (0.1, 0.2, 0.8)\); for \(\rho_2 = 0.5\), we set \(\rho_1 = (0.5, 0.7, 0.9)\) and \(\rho_3 = (0.5, 0.7, 0.9)\). Note that for \(K_1 = 100\), we only need to specify \(\rho_1 = (0, 0.1, 0.2, 0.5, 0.7, 0.9)\). In all cases, mul-
Figure 2.1: This figure shows the type I error rates and powers when the effects are homogeneous.

tivariate normal random samples (as summary statistics) were randomly generated with mean vector set to be the effect size vector $\mu$ and correlation matrix specified by $\rho_1$, $\rho_2$ and $\rho_3$. We repeat this simulation for 1000 times and set 0.05 as the significance level. The empirical sizes and powers were computed as the proportion of $p$-values less than 0.05 under the null and alternative models respectively.

Now we report the simulation results evaluating type I error rates and empirical powers.
Figure 2.2: This figure shows the type I error rates and powers when the effects are heterogeneous summarized in Figure 2.1 and 2.2. The type I error rates are always maintained and therefore our proposed methods are valid. The SUM test has the largest power when the effects are homogeneous and the signals are fully dense \((K_1 = 100)\), consistent with the theoretical results. However, SUM test has substantial power loss when the signals are sparse or the effects are in different directions. DSUM test has comparable powers as SUM when the effects are fully dense and homogeneous, and also loses power when
effects are sparse and heterogeneous. MinP test has the largest power when \( K_1 = 1 \), regardless of the correlation structures. Thus, MinP is best suited to detect super sparse signals. However, when the signals become denser, MinP loses power and therefore is not recommended for the detection of dense signals. WI test performs better than others for homogeneous effects settings when \( K_1 = 7, 25 \) and \( \rho_1 > \rho_3 \), but performs worse in other cases.

The variance component (VC) test generally performs poorly when the effects are homogeneous, but has the highest power when effects are heterogeneous when \( K_1 = 7, 25, 100 \) and \( \rho_1 > \rho_3 \). Wald test performs poorly when the signals are sparse, however performs the best for heterogeneous effects when \( K_1 = 7, 25 \) and \( \rho_3 > \rho_1 \) and has almost the same power as VC when \( K_1 = 100 \). Linear mixed model based methods (mixAda, mixFisher, mixTippett, mixSDWeight, mixVarWeight) although are less powerful than MinP but more powerful than other methods when \( K_1 = 1 \). mixVarWeight generally performs worse than other mixed model based methods. When the effects are homogeneous and \( K_1 = 7, 25 \), mixed model based methods are more powerful when \( \rho_3 > \rho_1 \) but are almost powerless when \( \rho_1 > \rho_3 \). When the effects are heterogeneous and \( K_1 = 7, 25 \), mixed model based methods are in general very powerful and robust to correlation structures, except that mixVarWeight is very sensitive to \( \rho_3 \) and could be powerless when \( \rho_3 > \rho_1 \). When \( K_1 = 100 \) and the effects are heterogeneous, mixed model based methods (except for mixVarWeight) are as powerful as VC and Wald. Therefore, mixAda, mixFisher, mixTippett and mixSDWeight are suited for both dense and sparse signals, and are relatively more robust to correlation structures and effect heterogeneity.

2.4 Global Lipids GWAS Data Reanalysis

We applied our methods to the global lipids GWAS data set which contains publicly available summary statistics assessing the associations between 2.6 millions directly genotyped or imputed SNPs and four lipids levels separately on more than 100,000 individuals of European ancestry (Teslovich et al., 2010). The original analysis conducted by Teslovich et al. (2010) found 95 genome-wide significant SNPs associated with at least one of the
<table>
<thead>
<tr>
<th>SNP</th>
<th>CHR</th>
<th>Gene</th>
<th>HDL Z-score</th>
<th>LDL Z-score</th>
<th>TC Z-score</th>
<th>TG Z-score</th>
<th>SUM P-value</th>
<th>VC P-value</th>
<th>mixFisher P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4238103</td>
<td>12</td>
<td>LARP4</td>
<td>0.22</td>
<td>-3.55</td>
<td>0.23</td>
<td>-0.63</td>
<td>3.1e-14</td>
<td>6.4e-27</td>
<td>2.0e-21</td>
</tr>
<tr>
<td>rs16959082</td>
<td>17</td>
<td>GAS7</td>
<td>-1.17</td>
<td>-3.96</td>
<td>-0.48</td>
<td>0.42</td>
<td>2.3e-14</td>
<td>1.1e-24</td>
<td>1.1e-19</td>
</tr>
<tr>
<td>rs6074916</td>
<td>20</td>
<td>MACROD2</td>
<td>0.89</td>
<td>1.59</td>
<td>-0.92</td>
<td>1.22</td>
<td>1.2e-13</td>
<td>2.9e-21</td>
<td>2.2e-19</td>
</tr>
<tr>
<td>rs7517253</td>
<td>1</td>
<td>LRRC8D</td>
<td>-4.59</td>
<td>-0.18</td>
<td>0.51</td>
<td>-0.01</td>
<td>3.8e-14</td>
<td>1.5e-14</td>
<td>2.0e-16</td>
</tr>
<tr>
<td>rs1209695</td>
<td>22</td>
<td>SEZ6L</td>
<td>0.53</td>
<td>1.65</td>
<td>-1.34</td>
<td>-0.26</td>
<td>1.1e-09</td>
<td>2.0e-19</td>
<td>2.2e-16</td>
</tr>
<tr>
<td>rs10418460</td>
<td>19</td>
<td>ZNF235</td>
<td>0.32</td>
<td>0.95</td>
<td>-1.72</td>
<td>0.08</td>
<td>1.8e-08</td>
<td>4.8e-17</td>
<td>6.9e-15</td>
</tr>
<tr>
<td>rs5167</td>
<td>19</td>
<td>APOC2</td>
<td>-4.81</td>
<td>-1.79</td>
<td>-4.11</td>
<td>-4.20</td>
<td>2.9e-12</td>
<td>2.5e-03</td>
<td>1.5e-13</td>
</tr>
<tr>
<td>rs3099840</td>
<td>6</td>
<td>HCP5</td>
<td>3.32</td>
<td>2.02</td>
<td>5.44</td>
<td>5.43</td>
<td>1.7e-05</td>
<td>2.5e-03</td>
<td>2.3e-13</td>
</tr>
<tr>
<td>rs12721444</td>
<td>13</td>
<td>TNFSF11</td>
<td>-0.78</td>
<td>-1.90</td>
<td>-0.12</td>
<td>-1.88</td>
<td>2.9e-11</td>
<td>2.6e-14</td>
<td>2.3e-13</td>
</tr>
<tr>
<td>rs3001673</td>
<td>1</td>
<td>PLD5</td>
<td>1.67</td>
<td>1.92</td>
<td>-0.52</td>
<td>-0.79</td>
<td>2.7e-09</td>
<td>8.8e-16</td>
<td>2.4e-13</td>
</tr>
<tr>
<td>rs751320</td>
<td>18</td>
<td>WDR7</td>
<td>0.32</td>
<td>2.22</td>
<td>-0.15</td>
<td>0.95</td>
<td>1.6e-09</td>
<td>5.4e-15</td>
<td>1.2e-12</td>
</tr>
<tr>
<td>rs874743</td>
<td>19</td>
<td>RELB</td>
<td>2.38</td>
<td>0.25</td>
<td>0.39</td>
<td>3.46</td>
<td>1.3e-12</td>
<td>1.6e-09</td>
<td>1.6e-12</td>
</tr>
<tr>
<td>rs6977264</td>
<td>7</td>
<td>hCG1745121</td>
<td>1.63</td>
<td>-2.53</td>
<td>0.34</td>
<td>-2.06</td>
<td>1.4e-08</td>
<td>7.0e-16</td>
<td>1.7e-12</td>
</tr>
<tr>
<td>rs474339</td>
<td>11</td>
<td>SIDT2</td>
<td>-2.91</td>
<td>-2.10</td>
<td>-3.97</td>
<td>-5.44</td>
<td>1.4e-11</td>
<td>1.9e-03</td>
<td>2.1e-12</td>
</tr>
<tr>
<td>rs3736120</td>
<td>11</td>
<td>PAFAH1B2</td>
<td>-3.14</td>
<td>-2.00</td>
<td>-3.92</td>
<td>-5.17</td>
<td>2.6e-11</td>
<td>2.5e-03</td>
<td>3.6e-12</td>
</tr>
<tr>
<td>rs11659343</td>
<td>18</td>
<td>ATP9B</td>
<td>1.96</td>
<td>0.55</td>
<td>-0.33</td>
<td>1.75</td>
<td>1.2e-10</td>
<td>2.1e-11</td>
<td>4.1e-12</td>
</tr>
<tr>
<td>rs664971</td>
<td>11</td>
<td>TAGLN</td>
<td>-3.08</td>
<td>-1.60</td>
<td>-3.43</td>
<td>-5.19</td>
<td>1.1e-11</td>
<td>8.7e-04</td>
<td>6.7e-12</td>
</tr>
<tr>
<td>rs2288798</td>
<td>5</td>
<td>SLIT3</td>
<td>-2.27</td>
<td>0.37</td>
<td>-2.32</td>
<td>1.84</td>
<td>1.8e-05</td>
<td>1.4e-13</td>
<td>7.9e-12</td>
</tr>
<tr>
<td>rs2280328</td>
<td>11</td>
<td>PEX16</td>
<td>-0.14</td>
<td>-1.91</td>
<td>0.96</td>
<td>0.55</td>
<td>6.4e-07</td>
<td>6.8e-15</td>
<td>8.5e-12</td>
</tr>
<tr>
<td>rs3130660</td>
<td>6</td>
<td>FLOT1</td>
<td>-2.99</td>
<td>-2.59</td>
<td>-4.85</td>
<td>-4.99</td>
<td>2.7e-08</td>
<td>4.5e-02</td>
<td>1.4e-11</td>
</tr>
</tbody>
</table>

Table 2.1: A subset of the top associated SNPs with lipids levels based on the p-values of mixFisher method. CHR refers to the chromosome number. SUM refers to the SUM test targeting for homogeneous effects. VC refers to the variance component test targeting for heterogeneous effects. And mixFisher refers to the Fisher’s method of combining two independent p-values, which is expected to be more robust than both SUM and VC. HDL, LDL, TC and TG refer to high and low density lipoprotein cholesterol, total cholesterol and triglycerides respectively.

Our methods have identified hundreds of novel genome-wide significant SNPs that were missed by the original analysis. We found a clustering pattern of the novel SNPs and those identified clusters are mostly located around known genes, which suggests plausible biological functions. For example, 17 novel intronic SNPs are from gene ABCA1 on chromosome 9. Actually, gene ABCA1 tagged by another SNP has already been reported to be negatively associated with high-density lipoprotein (HDL) cholesterol (Teslovich et al., 2010), here we found extra novel SNPs in this gene. In total, we found 210 unique genes associated with lipids levels across the whole genome. For illustration purpose,
Figure 2.3: This figure shows the novel SNPs identified by each method proposed in this paper. The Wald test was used as the comparison reference. Using the top-left panel as an example, we see that Wald test identified 780 novel SNPs and 89 of them were also found by the SUM test. Meanwhile, SUM test identified 4 novel SNPs that both Wald and DSUM were unable to discover.

we present the top 20 genes in Table 2.1 based on the $p$-values of mixFisher method. We found that SNP rs4238103 located in gene LARP4 (protein coding gene) on chromosome 12 was not identified by univariate GWAS approaches since none of the four Z-scores indicates genome-wide significance. Our SUM, VC and mixFisher tests yield $p$-values equal to $3.1 \times 10^{-14}$, $6.4 \times 10^{-27}$, and $2.0 \times 10^{-21}$ respectively. Furthermore, our mix-Fisher method can additionally find that SNP rs4238103 has both shared group effects
(\(p = 3.2 \times 10^{-14}\)) and individual effects (\(p = 1.3 \times 10^{-9}\)) on the four lipids levels. Another SNP rs3130660 in gene FLOT1 on chromosome 6 was found to have only shared group effects (\(p = 2.9 \times 10^{-8}\)) on the four lipids. On the contrary, SNP rs22889798 was found to have only individual heterogeneous effects (\(p = 1.4 \times 10^{-8}\)) on the four lipids. These examples illustrate that our methods can dramatically increase statistical powers to detect genetic variants with complicated effects on the lipids. Moreover, the association results can be well interpreted.

We also present Figure 2.3 to show the number of novel SNPs identified and co-identified by the proposed methods. We can see that Wald, mixFisher, mixSDWeight and mixAda identified 780, 535, 513 and 394 novel SNPs respectively and DSUM, VC, mixVarWeight, mixTippett and SUM identified 221, 209, 141, 129 and 93 novel SNPs respectively. Note that some of the novel SNPs might be discovered by more than one methods at the same time, for example, there are 118 novel SNPs identified by Wald, mixTippett and mixFisher. There are 96 novel SNPs detected by mixFisher, but not by Wald and mixTippett as shown in the top-right panel of Figure 2.3. Although no method is uniformly better than the others, however we can see that our mixed model based testing procedures perform well in both homogeneous and heterogeneous effects settings.

2.5 Discussion

In this paper, we propose several statistical testing procedures to detect genetic variants associated with multiple correlated phenotypes based on summary statistics. Our methods take the correlation structure among the multiple phenotypes explicitly into account and therefore can potentially increase statistical powers.

The proposed testing procedures have attractive advantages. First, our methods only need summary statistics to perform multiple phenotype association analysis. The proposed tests are very useful since more and more summary statistics are publicly available, while individual level data are not easily accessible due to administration issues. Second, our methods are computationally efficient since they all have analytic formulas to compute the \(p\)-values. This makes our methods considerably appealing in large-scale
genetic association studies where millions of genetic markers are assessed with multiple phenotypes, and hence computational efficiency is of top priority. Third, our methods can deal with mixed types of phenotypes, for example, binary or continuous phenotypes, two most common phenotypes in biomedical research. This is also appealing for practical uses since most joint modeling of mixed types of phenotypes using individual level data is always a statistical and computational challenge (Fitzmaurice and Laird, 1997). Fourth, our methods can also identify the sources contributing to the associations between a genetic variant and multiple phenotypes, and therefore can help us better interpret the association results. Fifth, the underlying genetic architecture varies from one gene to another across the whole genome. Our mixed model based tests can capture both homogeneous and heterogeneous effects, and are therefore robust to genetic effect patterns. An application to a global lipids GWAS data set identified hundreds of novel genetic variants that were missed by conventional single-trait methods, which demonstrates the usefulness of our methods in practice.

The proposed testing procedures are well suited for analyzing multiple phenotypes measured on the same study subjects. Extensions of our methods for non-overlapping samples is of future research interest. Biological interpretation of GWASs findings is always a challenging problem. Our methods shed new light on interpreting pleiotropic associations by decomposing the genetic effects into shared group and individual effects. However, future research is still needed to measure the strength of pleiotropy.
A Geometric Perspective on the Powers of Principal Component Association Tests

Zhonghua Liu and Xihong Lin
Department of Epidemiology and Biostatistics
Harvard T.H. Chan School of Public Health
3.1 Introduction

There is increasing interest to analyze multiple phenotypes jointly in genome-wide association studies (GWASs) to detect pleiotropic effects, and the power gain of joint analysis of multiple phenotypes has been demonstrated by several previous studies (Roy et al., 2003; Klei et al., 2008; Yang et al., 2010; Yang and Wang, 2012; Schifano et al., 2013; Stephens, 2013; Solovieff et al., 2013; Zhou and Stephens, 2014). Principal component analysis (PCA) is a popular method for dimension reduction which has been proposed for multiple phenotypes association studies (Zhang et al., 2012; Karasik et al., 2012; Suo et al., 2013; Aschard et al., 2014). The idea is to use the principal components of the original phenotypes as new composite scores for downstream genetic association testings.

However, there is no consensus on how to choose the principal components to maximize the powers and the existing ad hoc methods lack theoretical basis, leading to potential power loss if researchers apply them inappropriately to their own studies. Moreover, previous PCA methods usually require raw individual level continuous phenotype and genotype data, which might be unfeasible in some situations. It is still a statistical challenge on how to conduct multiple phenotype association studies without access to individual level data and how to integrate information from both dichotomous and continuous phenotype.

In this paper, we propose several principal component association tests (PCAT) for analyzing multiple phenotypes to detect underlying causal genetic variants with pleiotropic effects, applicable to both individual level data and summary statistics. Specifically, we apply orthogonal transformation to the summary statistics and obtain uncorrelated principal components which are therefore mutually independent under multivariate normal assumption. We further discuss why and when using only top few principal components for association testings might lose power substantially from a geometric perspective by introducing the concept of principal angles (PA), defined as the angles between the effect size vector and the eigenvectors of the correlation matrix. To increase statistical powers, we propose several PCAT methods to aggregate the information contained in all the principal components by combining either principal components or the corresponding
\( p \)-values. All of our methods have analytic formulas to compute \( p \)-values and therefore are computationally efficient even when the number of phenotypes is large. We also perform extensive simulation studies to compare the relative performances of the proposed methods in a wide range of settings. Lastly, we apply our methods to a global lipids level genome-wide association study data set and identify hundreds of novel genetic variants that have been missed by univariate analysis approaches. The results presented here can help better guide researchers to choose appropriate principal component based methods for association studies under various circumstances.

In section 3.2, we describe our PCAT methods and discuss their theoretical powers under various situations. In section 3.3, we perform eigen-analysis by explicitly deriving the closed form expressions for the eigenvalues and eigenvectors of typical correlation matrices and discuss how the correlation structures influence the statistical powers of PCAT methods. In section 3.4, we conduct extensive simulation studies to assess the empirical performances of our methods. In section 3.5, we apply PCAT procedures to a global lipids GWAS data set to illustrate the usefulness of our methods. And the article ends with a discussion in section 3.6.

3.2 Principal Component Association Tests

Single-trait genome-wide association studies have generated an output matrix of summary statistics with rows and columns representing genetic variants and multiple phenotypes respectively. And each element of this output matrix is the Wald-type \( Z \) testing statistic. Suppose that there are \( K \) correlated phenotypes denoted by \( Y = (Y_1, \ldots, Y_K)^T \), which are sometimes not easily accessible in many situations. The correlation among the \( K \) phenotypes induces the correlation among the corresponding testing statistics \( Z = (Z_1, \ldots, Z_K)^T \) for each genetic variant. The easily accessible summary statistics \( Z \) asymptotically follows a multivariate normal distribution with variance-covariance matrix equal to the correlation matrix of \( Y \) under the null. Although across the whole genome, either genotyped or imputed genetic variants could have different allele frequencies, their association testing statistics \( Z \) follow the same null distribution irrespective of
the allele frequencies. This serves as the basis for estimating the variance-covariance matrix of $Z$ using its sample version over all the null genetic markers after proper linkage disequilibrium (LD) pruning.

### 3.2.1 Single Principal Component Method

We assume that the variance-covariance matrix $\Sigma$ of $Z$ is known at this stage for the sake of simplicity. The mean vector of $Z$ under the alternative hypothesis is denoted as $\beta$ with elements equal to zeros for null genetic variants and non-zero for non-null genetic variants. Therefore, we have $Z = (Z_1, \ldots, Z_K)^T \sim N(\beta, \Sigma)$ for a particular genetic variant. Note that $\Sigma$ can be decomposed as $\Sigma = U \Lambda U^T = \sum_{k=1}^{K} \lambda_k u_k u_k^T$, where $\Lambda$ is a diagonal matrix $\text{diag}(\lambda_1, \ldots, \lambda_K)$ with elements representing eigenvalues and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$, and $U$ is the normalized orthogonal matrix with $k$th column $u_k$ representing the $k$th eigenvector associated with the $k$th largest eigenvalue. Actually, the eigenvectors are not unique by observing that both $u_k$ and $-u_k$ are valid candidates. This non-uniqueness could lead to substantial power loss for principal component based methods if we choose the wrong signs of the eigenvectors. In this article, we address this non-uniqueness by choosing the $k$th eigenvector such that $u_k^T \beta \geq 0$ (Note that common computing software might choose the signs of eigenvectors arbitrarily). Then the distribution of the $k$th principal component is given by

$$PC_k = u_k^T Z \sim N(u_k^T \beta, \lambda_k), 1 \leq k \leq K.$$  (3.1)

We now provide theoretical power analysis to investigate how each relevant parameter influences the power of single PC method. Using the fact that $||u_k||^2 = 1$, the non-centrality parameter ($ncp$) is

$$ncp = \frac{(u_k^T \beta)^2}{\lambda_k} = \frac{||\beta||^2 \cos^2(\theta_k)}{\lambda_k}.$$  

where $\theta_k \in [0, 90^\circ]$ is the $k$th principal angle (PA) defined as the angle between $\beta$ and the eigenvector $u_k$. An underlying constraint for the principal angles is that $\sum_{k=1}^{K} \cos^2(\theta_k) = 1$. If we use $PC_k$ as a testing statistic for the hypothesis: $H_0: \beta = 0$ versus $H_a: \beta \neq 0$, then
its theoretical power is

\[
\text{Power} = \Phi(Z_{\frac{\alpha}{2}} + \sqrt{ncp}) + \Phi(Z_{\frac{\alpha}{2}} - \sqrt{ncp}),
\]

where \(\Phi(\cdot)\) is the cumulative distribution function of a standard normal random variable, and \(Z_{\frac{\alpha}{2}}\) is its lower \(\frac{\alpha}{2}\) percentile with significance level \(\alpha\). For testing statistic \(PC_k\), it could be powerful if \(\theta_k = 0\) but powerless if \(\theta_k = 90^\circ\). This geometric perspective clearly explains why the first PC (and actually any other PCs) might be powerless to detect the signals in certain situations. Unfortunately, researchers usually have no prior information about the principal angles in practice and therefore are not warranted to have good powers of using any single PC for association testings. Therefore more robust and powerful methods are desired. To understand the best and worst situations of a testing procedure, we now define two terms, the most favorable alternative (MFA) and the least favorable alternative (LFA) as the alternative distributions under which the powers of the testing statistic are maximized and minimized respectively.

### 3.2.2 PCMinP

Instead of using any single PC with the potential danger of power loss, we propose to use the minimum of the principal component \(p\)-values as a testing statistic given by

\[
\text{PCMinP} = \min_{1 \leq k \leq K} p_k,
\]

where \(p_k\) is the \(p\)-value based on \(PC_k\). Actually, PCMinP is equivalent to \(\sup_{1 \leq k \leq K} \frac{|PC_k|}{\sqrt{\lambda_k}}\).

Note that PCMinP follows a Beta distribution \(\text{Beta}(\alpha = 1, \beta = K)\) under the null hypothesis, so its \(p\) value can be easily computed as

\[
P_{\text{PCMinP}} = 1 - (1 - \text{PCMinP})^K. \tag{3.2}
\]

\(P_{\text{PCMinP}}\) can be interpreted as the estimated probability of observing at least one principal component \(p\)-value less than PCMinP under the null hypothesis. It is closely related to Sidak adjusted \(p\)-value for multiple independent tests (Sidak, 1967; Wright, 1992). To assess the power of PCMinP under the alternative hypothesis, we obtained the probability
density function $g(p)$ of $p_k$ given by

$$g_{p_k}(p; \delta = ncp_k) = \frac{f(\chi^2_{1,1-p}; df = 1, \delta = ncp_k)}{f(\chi^2_{1,1-p}; df = 1, \delta = 0)}, \quad ncp_k = (u_k^T \beta)^2 / \lambda_k,$$

where $f(\cdot)$ represents the probability density function of a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\delta$ and $\chi^2_{1,1-p}$ represents the $(1 - p)th$ percentile of a central chi-squared distribution with one degree of freedom.

At significance level $\alpha$, the power of PCMinP is given by

$$\text{Power} = 1 - Pr\left(p_1 \geq 1 - (1 - \alpha)^{1/K}, \ldots, p_K \geq 1 - (1 - \alpha)^{1/K} | H_a\right) = 1 - \prod_{k=1}^{K} \int_{1-(1-\alpha)^{1/K}}^{1} g_{p_k}(p; \delta = ncp_k) dp.$$

We observe that PCMinP will be powerful if a small number of PCs can capture all the signals and especially powerful if the last PC captures all the signals. However, PCMinP will lose power if a large number of PCs have weak or modest signals.

### 3.2.3 PCFisher

PCMinP aims to pick the best PC direction and discard the other PCs, thus it loses power when a large number of PCs contain signal information. We hereby propose to combine all the $K$ independent principal component $p$-values using Fisher’s method (Fisher, 1932) with its null distribution given by

$$\text{PCFisher} = -2 \sum_{k=1}^{K} \log(p_k) \sim \chi^2_{2K},$$

where $p_k$ is the $p$-value for the $k$th principal component. PCFisher aggregates all the information contained in each PC. It assigns larger weights to PCs with smaller eigenvalues and therefore achieves its maximal power when the last PC captures all the signals.

In general, there is no uniformly most powerful (UMP) tests for composite alternatives with multiple parameters (Cox and Hinkley, 1974), therefore Fisher’s method is not UMP. Nevertheless, Fisher’s method is asymptotically optimal among essentially all methods of combining independent tests linearly according to Bahadur relative efficiency (Bahadur, 1971).

### 3.2.4 PCLC

In addition to the \( p \)-value based combination methods, we can also consider combining PCs directly. We propose a linear combination of the PCs (PCLC) using inverse variance weighting scheme with null distribution given by

\[
PCLC = \sum_{k=1}^{K} \frac{PC_k}{\lambda_k} \sim N(0, \sum_{k=1}^{K} \lambda_k^{-1}).
\]

Under the alternative hypothesis, its non-centrality parameter is

\[
ncp = \left( \sum_{k=1}^{K} \lambda_k^{-1} u_k^T \beta \right)^2 = \frac{||\beta||^2 \left( \sum_{k=1}^{K} \lambda_k^{-1} \cos(\theta_k) \right)^2}{\sum_{k=1}^{K} \lambda_k^{-1}}.
\]

We now derive the optimal power of PCLC for any fixed \( ||\beta||^2 \) by solving the following constrained optimization problem

\[
\begin{align*}
\max_{\cos(\theta_k)} & \quad \sum_{k=1}^{K} \lambda_k^{-1} \cos(\theta_k) \\
\text{s.t.} & \quad \sum_{k=1}^{K} \cos^2(\theta_k) = 1 \\
& \quad \cos(\theta_k) \geq 0, \quad \forall k \in \{1, \ldots, K\}.
\end{align*}
\]

(3.3)

A direct application of Lagrange multiplier method gives the optimal solution

\[
\cos(\theta_k) = \frac{\lambda_k^{-1}}{\sqrt{\sum_{k=1}^{K} \lambda_k^{-2}}}, \quad k = 1, \ldots, K.
\]

Therefore, the optimal power occurs when the principal angles \( \theta_k \) are in the above specified decreasing order. In other words, this requires that the first PC has the largest angle with \( \beta \) and the last PC has the smallest angle with \( \beta \). And any non-null distributions \( N(\beta, \Sigma) \) that give rise to such optimal principal angles are the most favorable alternative (MFA) distributions. In contrast, if

\[
\sin(\theta_k) = \frac{\lambda_k^{-1}}{\sqrt{\sum_{k=1}^{K} \lambda_k^{-2}}}, \quad k = 1, \ldots, K,
\]

then PCLC has its lowest power.
3.2.5 PCBall

We could also combine principal components using the Euclidean distance given by

\[ \text{PCBall} = \sum_{k=1}^{K} PC_k^2. \]

PCBall follows a chi-square mixture distribution \( \sum_j \lambda_j \chi_{j}^2 \), where \( \lambda_j \) are the eigenvalues of \( \Sigma \). And the \( p \)-value can be computed using either exact method (Davies, 1980) or moment matching method (Liu et al., 2009). We can also show that \( \sum_{k=1}^{K} PC_k^2 \) is equivalent to \( \sum_{k=1}^{K} Z_k^2 \) using the isometric property of orthogonal matrices, that is \( \sum_{k=1}^{K} Z_k^2 = Z^T U \Sigma^{-1} U^T Z = \sum_{k=1}^{K} PC_k^2 \), where \( U \) is the eigenvector matrix of \( \Sigma \). At significance level \( \alpha \), we reject the null hypothesis \( H_0: \beta = 0 \) if \( P(\sum_{k=1}^{K} PC_k^2 > C_\alpha; H_0) = \alpha \). Thus, the acceptance region of PCBall is a \( K \)-dimensional ball with radius equal to \( \sqrt{C_\alpha} \). Under the alternative hypothesis, the probability mass outside of this acceptance ball is the statistical power which depends on the mean (location) and the correlation matrix (shape) of the alternative distribution.

3.2.6 PCSS

The \( K \)-dimensional ball might not be the optimal acceptance region in some situations since PCBall doesn’t take the variance of each PC explicitly into account. We now propose a combination method using the sum of squared scaled PCs (PCSS) with null distribution given by

\[ \text{PCSS} = \sum_{k=1}^{K} \frac{PC_k^2}{\lambda_k} \sim \chi_k^2. \]

Actually, PCSS is equivalent to the Wald-type test \( Z^T \Sigma^{-1} Z \) by observing that \( Z^T \Sigma^{-1} Z = (U^T Z)^T \Lambda^{-1} (U^T Z) = \sum_{k=1}^{K} PC_k^2 / \lambda_k \), where \( \Sigma^{-1} = U \Lambda^{-1} U^T \). At significance level \( \alpha \), its acceptance region is determined by \( \sum_{k=1}^{K} PC_k^2 / \lambda_k \leq C_\alpha \), which is a \( K \)-dimensional ellipsoid, where \( C_\alpha \) is the \( 1 - \alpha \) percentile of \( \chi_k^2 \). Under the alternative, the distribution of PCSS is a non-central chi-squared distribution with non-centrality parameter \( ncp = \sum_{k=1}^{K} ||\beta||^2 \cos^2(\theta_k) / \lambda_k \). It is of interest to understand when PCSS achieves its maximal power for any given \( ||\beta||^2 \). We formulate this problem as the following constrained
optimization problem

\[
\begin{align*}
\max_{\cos^2(\theta_k)} & \quad \sum_{k=1}^{K} \lambda_k^{-1} \cos^2(\theta_k) \\
\text{s.t.} & \quad \sum_{k=1}^{K} \cos^2(\theta_k) = 1, \\
& \quad \cos(\theta_k) \geq 0, \quad \forall k \in \{1, \ldots, K\}.
\end{align*}
\] (3.4)

Using standard linear programming technique, we obtain that the solution is achieved when \(\cos(\theta_K) = 1\), which is one of the \(K\) vertices of the simplex given by the constraints in equation (3.4). In addition, its minimal power occurs when \(\cos(\theta_1) = 1\). Therefore, the most favorable alternatives for PCSS are the distributions whose last principal angles are zeros, and the least favorable alternatives are those distributions whose first principal angles are zeros.

### 3.2.7 PCWSS

PCLC suffers from the problem that different PCs might have different signs and thus the association signals might be diluted after summation. To overcome this drawback, we use inverse eigenvalue weighted sum squares of PCs (PCWSS) to combine principal components given by

\[
\text{PCWSS} = \sum_{k=1}^{K} \left( \frac{PC_k}{\lambda_k} \right)^2.
\]

It can be shown that PCWSS is equivalent to the variance component score testing statistic under a random effect model given by \(VC = Z^T \Sigma^{-1} \Sigma^{-1} Z\) (Huang and Lin, 2013). To see this connection, note that \(Z^T \Sigma^{-1} \Sigma^{-1} Z = (U^T Z)^T \Lambda^{-1} \Lambda^{-1} (U^T Z) = \sum_{k=1}^{K} \left( \frac{PC_k}{\lambda_k} \right)^2\). PCWSS follows a mixture of chi-square distributions \(\sum_{k=1}^{K} \lambda_k^{-1} \chi^2_{1k}\), where \(\lambda_k\) are the eigenvalues of \(\Sigma\). The \(p\)-value of PCWSS can be computed using either exact method (Davies, 1980) or moment matching method (Liu et al., 2009). Geometrically, the acceptance region of PCWSS is also a \(K\)-dimensional ellipsoid. So the most favorable alternatives are those whose last principal angles are zeros.

### 3.2.8 Graphical Demonstration of Rejection Boundaries

We now present a hypothetical example to illustrate the differences and connections among all the principal component based methods from a geometric perspective. Sup-
Figure 3.1: This figure illustrates the acceptance regions of PC1, PC2, PCMinP and PC-Fisher methods. The red solid lines represent the boundaries that separate the acceptance and rejection regions. For each method, we generated 10000 random samples from two standard bi-variate normal distributions with mean $(2.6, 2.6)$ and $(-2.6, 2.6)$ respectively, with the same correlation coefficient 0.8. The blue dashed lines are marked at $-1.96$ and $1.96$ for the horizontal and vertical axis representing the critical values of univariate tests at significance level 0.05.
Figure 3.2: This figure illustrates the acceptance regions of PCLC, PCBall, PCSS and PCWSS methods. The red solid lines represent the boundaries that separate the acceptance and rejection regions. For each method, we generated 10000 random samples from two standard bi-variate normal distributions with mean $(2.6, 2.6)$ and $(-2.6, 2.6)$ respectively, with the same correlation coefficient 0.8. The blue dashed lines are marked at $-1.96$ and $1.96$ for the horizontal and vertical axis representing the critical values of univariate tests at significance level 0.05.
pose that \((Z_1, Z_2)^T\) follows a bivariate normal distribution \(N(\beta, \Sigma)\), where \(\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\).

We are interested in testing whether \(H_0: \beta = 0\). First, the eigenvalues of \(\Sigma\) are \(\lambda_1 = 1.8\) and \(\lambda_2 = 0.2\), and the eigenvectors are given by \(u_1^T = (1/\sqrt{2}, 1/\sqrt{2})\) and \(u_2^T = (-1/\sqrt{2}, 1/\sqrt{2})\). So the two PCs are \(PC_1 = Z_1/\sqrt{2} + Z_2/\sqrt{2}\) and \(PC_2 = -Z_1/\sqrt{2} + Z_2/\sqrt{2}\) respectively.

We first determine the rejection boundaries under the null at significance level 0.05 for the proposed principal component association tests including \(PC_1\), \(PC_2\), \(PC\text{MinP}\), \(PC\text{Fisher}\), \(PCLC\), \(PC\text{Ball}\), \(PC\text{SS}\) and \(PC\text{WSS}\). The rejection boundaries of \(PC_1\) and \(PC_2\) are determined by \(PC_1 = \pm 1.96 \times \sqrt{1.8} \approx \pm 2.63\) and \(PC_2 = \pm 1.96 \times \sqrt{0.2} \approx \pm 0.88\). Geometrically, the rejection boundaries for \(PC_1\) and \(PC_2\) are both two parallel straight lines in the two dimensional space. As for \(PC\text{MinP}\), its rejection boundary is determined jointly by \(PC_1 = \pm \Phi^{-1}((1+\sqrt{0.95})/2) \times \sqrt{1.8} = \pm 3.0\) and \(PC_2 = \pm \Phi^{-1}((1+\sqrt{0.95})/2) \times \sqrt{0.2} = \pm 1.0\), which is a rectangle. As for \(PC\text{Fisher}\), its rejection boundary is determined by the following \(\log(\Phi(-|PC_1|/\sqrt{1.8})) + \log(\Phi(-|PC_2|/\sqrt{0.2})) = -F_{\chi^2_4}^{-1}(0.95)/2\), where \(F_{\chi^2_4}^{-1}(0.95)\) gives the 95% percentile of \(\chi^2_4\). The rejection boundaries of \(PC_1\), \(PC_2\), \(PC\text{MinP}\) and \(PC\text{Fisher}\) are depicted in Figure 3.1.

We now discuss the other four principal component combination methods. As for \(PCLC\), its rejection boundary is determined by \(P(|PC_1/1.8 + PC_2/0.2| > 1.96\sqrt{1/1.8 + 1/0.2}; H_0) = 0.05\), which are two straight lines : \(-3.14Z_1 + 3.93Z_2 = \pm 4.62\).

The rejection boundary of \(PC\text{Ball}\) is a circle with its center located in the origin and radius equal to 7.2 determined by \(P(Z_1^2 + Z_2^2 > 7.2; H_0) = 0.05\). And for \(PC\text{SS}\) and \(PC\text{WSS}\), their rejection boundaries are the ellipses determined by \(PC_1^2/1.8 + PC_2^2/0.2 = 5.99\) and \(PC_1^2/1.8^2 + PC_2^2/0.2^2 = 20\) respectively. We also plot the rejection boundaries for \(PCLC\), \(PC\text{Ball}\), \(PC\text{SS}\) and \(PC\text{WSS}\) in Figure 3.2.

In both Figure 3.1 and 3.2, we also add 10000 random points randomly sampled from the following two alternative bivariate normal distributions with mean vectors \(\beta_1^T = (2.6, 2.6)\) and \(\beta_2^T = (-2.6, 2.6)\), with the same covariance matrix as above. The proportions of data points outside of the rejection boundaries are empirical powers. Both figures show that all the principal component association tests (PCAT) favors the case when the last principal angle is zero (MFA) and all lose substantial powers when the first principal angle is zero (LFA). For MFA, \(PC_2\) performs better than all the rest methods. \(PCLC\), \(PC\text{WSS}\), \(PC\text{Fisher}\),
PCSS and PCMinP will also have good powers since they all assign larger weights for $PC_2$, while $PC_1$ is powerless since the first principal angle is $90^\circ$. Because PCBall weights the two PCs equally, so PCball is not powerful in this case. However, for LFA, $PC_1$ has the largest power than all the others. PCBall also performs much better than PCWSS, PCFisher and PCSS because it has shorter radius along the first PC direction. PCMinP has modest power since it also doesn’t favor $PC_1$ because its rejection boundary has longer edge along the $PC_1$ direction. PCLC and $PC_2$ are almost powerless in this case.

### 3.3 Eigen-analysis of Correlation Matrices

The purpose of eigen-analysis is to understand how the correlation structure $\Sigma$ affects its eigenvalues and eigenvectors, and eventually affects the powers of PCATs. Without loss of generality, we restrict our eigen-analysis to structured correlation matrices. Suppose that $K_1$ out of $K$ phenotypes are associated with the genetic variant, and $K_0$ of them are not. To investigate how signal sparsity affects statistical powers, we let $K_1 = K^{1-\gamma}$, where $\gamma \in [0, 1]$ is the sparsity parameter. The signals are sparse if $\gamma \in (\frac{1}{2}, 1]$ and dense otherwise. The mean vector of $Z$ under the alternatives is $\beta$, which contains $K_1$ non-zero values and $K_0$ zeros, denoted as $\beta^T = (\beta_1, \ldots, \beta_{K_1}, 0, \ldots, 0)$. The variance-covariance matrix of $Z$ is denoted as

$$
\Sigma = \begin{pmatrix}
\Sigma_1 & \Sigma_2 \\
\Sigma_2 & \Sigma_3
\end{pmatrix}
$$

where $\Sigma_1 = (1 - \rho_1)I_{K_1} + \rho_1 J_{K_1K_1}$, $\Sigma_2 = \rho_2 J_{K_1K_0}$ and $\Sigma_3 = (1 - \rho_3)I_{K_0} + \rho_3 J_{K_0K_0}$. Denote $\nu_\Sigma(\lambda_k)$ as the algebraic multiplicity of eigenvalue $\lambda_k$. If $\nu_\Sigma(\lambda_k) = 1$, then we say that the eigenvector associated with $\lambda_k$ forms an eigenline. And if $\nu_\Sigma(\lambda_k) > 1$, then the eigenvectors associated with $\lambda_k$ span into an eigenspace with dimension equal to $\nu_\Sigma(\lambda_k)$ (Serre, 2010). We also have to ensure the correlation matrices to be positive definite by requiring that all the eigenvalues are positive.
3.3.1 Exchangeable Correlation Matrices

If $\rho_1 = \rho_2 = \rho_3 = \rho$, then we can easily obtain the closed-form expressions for the unsorted eigenvalues

$$\lambda_1 = (K - 1)\rho + 1, \quad \lambda_k = 1 - \rho, \quad k = 2, \ldots, K.$$ 

If $0 < \rho < 1$, then $(K - 1)\rho + 1 > 1 - \rho > 0$. The first eigenvector associated with the largest eigenvalue $(K - 1)\rho + 1$ is $u_1^T = (\frac{1}{\sqrt{K}}, \frac{1}{\sqrt{K}}, \ldots, \frac{1}{\sqrt{K}})$. And the eigenspace associated with eigenvalue $1 - \rho$ is

$$E_{\lambda_1=1-\rho} = \{ u \in R^K : \sum_{k=1}^{K} u_k = 0 \}.$$ 

The dimension of $E_{\lambda_1=1-\rho}$ is $K - 1$, so there are infinitely many possible choices for the $K - 1$ eigenvectors when $K \geq 3$. The distribution of $PC_1$ is $N(\sum_{k=1}^{K} \beta_k \sqrt{K}, (K - 1)\rho + 1)$ with non-centrality parameter given by

$$ncp = \left(\frac{||\beta|| \cos(\theta_1)}{(K - 1)\rho + 1} \right)^2 \leq \frac{K^{1-\gamma}(\max_k |\beta_k|)^2}{K\rho + 1 - \rho} \leq \frac{(\max_k |\beta_k|)^2}{K^{\gamma} \rho + (1 - \rho)K^{\gamma-1}} \downarrow 0, \quad K \rightarrow +\infty, \quad (3.5)$$

assuming $\max_k |\beta_k| < \infty$ and $\gamma \in (0, 1]$. This says that $ncp$ will converge to 0 at the rate of $K^{-\alpha}$ as long as the signals are not fully dense. So the power of $PC_1$ decreases in $K$ and might be powerless when $K$ is large. However, if the signals are fully dense ($\gamma = 0$), then we have

$$\frac{(\min_k |\beta_k|)^2 \cos^2(\theta_1)}{\rho} \leq ncp \leq \frac{(\max_k |\beta_k|)^2 \cos^2(\theta_1)}{\rho}, \quad K \rightarrow +\infty.$$ 

It is therefore advised not to use $PC_1$ for association studies in the high dimensional settings if the signals are not fully dense. However, if the signals are fully dense, then $PC_1$ could be used to detect signals in high dimensional situations with good power. In both cases, smaller correlation leads to larger power. If the effects are homogeneous, then the first PC captures all the information and the other PCs are orthogonal to $\beta$ and hence contain no signal information.

If $\frac{1}{K-1} \leq \rho < 0$ , then $1 - \rho > (K - 1)\rho + 1 > 0$. So the largest eigenvalue is $1 - \rho$ and the associated eigenspace $E_{\lambda_1=1-\rho}$ has dimension $K - 1$. Consequently, there are infinitely many possible choices for the first $K - 1$ principal components when $K \geq 3$. It’s powerless to detect any homogeneous effects using any of the first $K - 1$ PCs because the effect size
vector $\beta$ is orthogonal to the whole eigenspace $E_{\lambda=1-\rho}$. For heterogeneous effects, the distribution of $PC_1$ is $N(u_1^T \beta, 1 - \rho)$ with non-centrality parameter given by

$$ncp = \frac{(||\beta|| \cos(\theta_1))^2}{1 - \rho} = \frac{\cos^2(\theta_1) \sum_{k=1}^{K_1-\gamma} \beta_k^2}{1 - \rho} \uparrow +\infty, \ K \rightarrow +\infty,$$

assuming $\gamma \in [0,1)$ and $\cos(\theta_1) \neq 0$. Therefore, stronger correlation and higher dimension leads to higher statistical power using $PC_1$.

### 3.3.2 Block Diagonal Exchangeable Correlation Matrices

If $\rho_2 = 0$, then the signals are uncorrelated with noises. The unsorted eigenvalues of $\Sigma$ and their algebraic multiplicities are given by

$$\lambda_1 = 1 + (K_1 - 1)\rho_1, \nu(\lambda_1) = 1; \lambda_2 = 1 - \rho_1, \nu(\lambda_2) = K_1 - 1;$$
$$\lambda_3 = 1 + (K_0 - 1)\rho_3, \nu(\lambda_3) = 1; \lambda_4 = 1 - \rho_3, \nu(\lambda_4) = K_0 - 1.$$

By solving the homogeneous linear system $(\Sigma - \lambda I)u = 0$, the signal eigenspaces are

$$E_{\lambda_1} = \{u \in R^K : u^T = t(1/\sqrt{K_1}, \ldots, 1/\sqrt{K_1}, 0, \ldots, 0), t \in R\},$$
$$E_{\lambda_2} = \{u \in R^K : \sum_{k=1}^{K_1} u_k = 0, u_{K_1+1} = \cdots = u_K = 0\},$$

and the noise eigenspaces are

$$E_{\lambda_3} = \{u \in R^K : u^T = t(0, \ldots, 0, 1/\sqrt{K_0}, \ldots, 1/\sqrt{K_0}), t \in R\},$$
$$E_{\lambda_4} = \{u \in R^K : \sum_{k=K_1+1}^K u_k = 0, u_1 = \cdots = u_{K_1} = 0\}.$$

Since the correlation matrix has to be positive definite, so $\rho_1 > -1/(K_1 - 1)$ and $\rho_3 > -1/(K_0 - 1)$. To be general, we assume that the four eigenvalues take distinct values.

1. If $\rho_1 > 0, 1 + (K_1 - 1)\rho_1 > 1 - \rho_3$ and $(K_1 - 1)\rho_1 > (K_0 - 1)\rho_3$, then $1 + (K_1 - 1)\rho_1$ is the largest eigenvalue with associated eigenspace $E_{\lambda_1}$. This situation occurs when the signals are very dense and the correlations among the signals are also strong. The
non-centrality parameter of $PC_1$ is $ncp = \frac{(\sum_{k=1}^{K_1} \beta_k)^2}{K_1(1+(K_1-1)\rho_1)}$. So the first PC is powerful if the effects are in the same direction but might be powerless if the effects are in different directions. In addition, smaller $\rho_1$ contributes to higher power.

2. If $\frac{-1}{K_1-1} \leq \rho_1 < 0$, $\rho_1 < \rho_3$ and $1 - \rho_1 > 1 + (K_0 - 1)\rho_3$, then the largest eigenvalue is $1 - \rho_1$ with eigenspace $E_{\lambda_2}$. If $K_1 \geq 3$, then there are infinitely many possible choices of eigenvectors for $E_{\lambda_2}$. For any chosen eigenvector, the entries for the noises are zeros and the sum of the entries for signals is also zero. It’s therefore not powerful to use this PC for homogeneous effects, but might be powerful for heterogeneous effects.

3. If $\rho_3 > 0$, $(K_0 - 1)\rho_3 > (K_1 - 1)\rho_1$ and $(K_0 - 1)\rho_3 > -\rho_1$, then the largest eigenvalue is $\lambda_3$ with associated eigenspace $E_{\lambda_3}$. This could happen when the signals are sparse and the correlations among the noises are strong. Since $\beta$ is orthogonal to $E_{\lambda_3}$, so this PC is always powerless.

4. If $\frac{-1}{K_0-1} \leq \rho_3 < 0$, $\rho_3 < \rho_1$ and $-\rho_3 > (K_1 - 1)\rho_1$, then the largest eigenvalue is $1 - \rho_3$ with eigenspace $E_{\lambda_4}$. The loading coefficients for the signals of any eigenvector of $E_{\lambda_4}$ are all zeros. Thus, the noises will completely dominate $PC_1$, and hence it’s powerless to detect any association signals.

### 3.3.3 Block Exchangeable Correlation Matrices

In practical settings, the inter-correlation $\rho_2$ between signals and noises is not zero. It’s interesting to know how nonzero $\rho_2$ affects the eigenvalues and eigenvectors. For $\rho_2 \neq 0$, the unsorted eigenvalues and their algebraic multiplicities (see appendix for derivations) are given by

\[
\begin{align*}
\lambda_1 &= \frac{1}{2} \left( -g + \sqrt{g^2 - 4h} \right), \quad \nu(\lambda_1) = 1; \\
\lambda_2 &= \frac{1}{2} \left( -g - \sqrt{g^2 - 4h} \right), \quad \nu(\lambda_2) = 1; \\
\lambda_3 &= 1 - \rho_1, \quad \nu(\lambda_3) = K_1 - 1; \\
\lambda_4 &= 1 - \rho_3, \quad \nu(\lambda_4) = K_0 - 1; \\
g &= -[2 + (K_1 - 1)\rho_1 + (K_0 - 1)\rho_3], \\
h &= 1 - \rho_3 + (K_1 - 1)\rho_1(1 - \rho_3) + K_0\rho_3 + K_0\rho_3(K_1 - 1)\rho_1 - \rho_2^2 K_1 K_0.
\end{align*}
\]
The eigenspaces associated with $\lambda_3$ and $\lambda_4$ usually have dimensions larger than 2 when $K_1 \geq 3$ and $K_0 \geq 3$, leading to infinitely many possible choices of the eigenvectors for the corresponding eigenspaces. Practically speaking, we are more interested in the other two well-specified eigenlines $E_{\lambda_1}$ and $E_{\lambda_2}$.

If $(K_1 - 1)\rho_1 < (K_0 - 1)\rho_3$, then the signals are stronger than the noises. The associated eigenvector for $E_{\lambda_1}$ is represented as

$$ u_{11}^2 = \cdots = u_{1K_1}^2 = \frac{[\lambda_1 - 1 - (K_0 - 1)\rho_3]^2}{K_1[\lambda_1 - 1 - (K_0 - 1)\rho_3]^2 + K_0K_1^2\rho_2^2} $$

$$ u_{1(K_1+1)}^2 = \cdots = u_{1K}^2 = \frac{K_0\rho_2^2}{K_1K_0\rho_2^2 + [\lambda_1 - 1 - (K_1 - 1)\rho_1]^2} $$

Note that the entries of $u_1$ are non-zeros, indicating that this eigenvector lies in between the signal and noise spaces. When $|\rho_2|$ increases, this eigenvector tends to depart from the signal space by increasing the weights for noises. Meanwhile, we have $\frac{\partial \lambda_1}{\partial \rho_2} = \frac{2\rho_2K_1K_0}{\sqrt{g^2-4h}}$, so $\lambda_1$ increases in $|\rho_2|$. Therefore, the $ncp = ||\beta||^2 \cos^2(\theta_1)/\lambda_1$ of using this PC will decrease. Hence larger $|\rho_2|$ tends to bring in more noises and dilute the signals.

However, if $(K_1 - 1)\rho_1 < (K_0 - 1)\rho_3$, then the signals are weaker than the noises. The associated eigenvector for $E_{\lambda_1}$ is

$$ u_{11}^2 = \cdots = u_{1K_1}^2 = \frac{K_0\rho_2^2}{K_1K_0\rho_2^2 + [\lambda_1 - 1 - (K_1 - 1)\rho_1]^2} $$

$$ u_{1(K_1+1)}^2 = \cdots = u_{1K}^2 = \frac{[\lambda_1 - 1 - (K_1 - 1)\rho_1]^2}{K_0(K_1K_0\rho_2^2 + [\lambda_1 - 1 - (K_1 - 1)\rho_1]^2)} $$

In this case, the increased value of $|\rho_2|$ might potentially help increase statistical power. Geometrically, by increasing $|\rho_2|$, this eigenvector tends to move towards the signal space by assigning more non-zero weights for the signals. Therefore, the numerator of the $ncp$ increases. However, as shown above, the eigenvalue $\lambda_1$ also increases in $|\rho_2|$. Overall, the $ncp$ might or might not increase when both the numerator and denominator of $ncp$ increase. Some numerical studies find that the power generally increases slightly for small $|\rho_2|$ and then might decease a little bit or stay almost constant.

For eigenspace $E_{\lambda_2}$, its eigenvector $u_2$ has the same expression as the one for eigenspace
$E_{\lambda_1}$ except replacing $\lambda_1$ by $\lambda_2$. When the signals are stronger than the noises, $u_2$ lies in the noise space when $\rho_2 = 0$, simply because $u_1$ and $u_2$ are orthogonal. Therefore, increasing $|\rho_2|$ will pull this eigenvector towards the signal space. Meanwhile, $\lambda_2$ is decreasing in $|\rho_2|$ by noting that $\frac{\partial \lambda_1}{\partial \rho_2} = \frac{-2\rho_2 K_1 K_0}{\sqrt{\rho_2^2 - 4h}}$. Therefore the power of this PC increases in $|\rho_2|$. When the signals are weaker than the noises, then increased value of $|\rho_2|$ will pull this eigenvector towards noise space. Meanwhile, the eigenvalue $\lambda_2$ still decreases in $|\rho_2|$. So the numerator of the $ncp$ decreases but its denominator increases when $|\rho_2|$ increases, so $ncp$ might or might not increase. Some numerical studies find that the power might in general increase slightly.

### 3.4 Simulation studies

We perform simulation studies to assess the type I error rates and powers of PCAT methods in a wide range of situations. In the simulation studies, we focus on how the dimension, signal sparsity, effect homogeneity and correlation structures all together influence the powers.

#### 3.4.1 Simulation Settings

We set the significance level to be $\alpha = 0.05$, and the empirical type I error rates and powers are calculated as the proportions of $p$-values that are less than $\alpha$ among 1000 replications under the null and alternatives respectively. The total number of phenotypes is set to be $K = 40$ and the numbers of true associations are set to be $K_1 = (1, 5, 15, 40)$ with $K_1 = (1, 5)$ represents sparse regime and $K_1 = (15, 40)$ represents dense regime. The correlation structures are set to be block matrices described previously in section 3.3. We first set $\rho_2 = (0, 0.1, 0.5)$ which represents three typical situations where signals and noises are uncorrelated, weakly and moderately correlated. For $\rho_2 = 0$, we set $\rho_1 = (0, 0.2, 0.8)$ and $\rho_3 = (0, 0.2, 0.8)$. As for $\rho_2 = 0.1$, we set $\rho_1 = (0.1, 0.2, 0.8)$ and $\rho_3 = (0.1, 0.2, 0.8)$; for $\rho_2 = 0.5$, we set $\rho_1 = (0.5, 0.7, 0.9)$ and $\rho_3 = (0.5, 0.7, 0.9)$. Note that when $K_1 = 40$, we only need to consider $\rho_1 = (0, 0.1, 0.2, 0.5, 0.7, 0.9)$. And for $K_1 = 1$, we only need to consider $\rho_2$ and $\rho_3$. Next, for homogeneous genetic effects, we set $(\beta_1, \ldots, \beta_{K_1}, 0, \ldots, 0) =$
(1, ..., 1, 0, ..., 0) × c where the scaling factor c ∈ [0, 4]. And for heterogeneous genetic effects, we set half of the K₁ effects to be negative and the other half to be positive. In the case K₁ is odd, we set one more positive component, for example, when K₁ = 5, we set (β₁, ..., β₅) = (1, 1, 1, −1, −1) × c where c ∈ [0, 4]. We implement all the methods introduced in this paper and also include MinP defined as \( \min_{1 \leq k \leq 40} p_k \) for comparison purpose, where \( p_k \) is the univariate p-value for the kth phenotype.

3.4.2 Simulation Results

Simulation results show that the type I error rates of our methods are all well-maintained. We now present the empirical power comparisons for homogeneous and heterogeneous effects respectively.

**Homogeneous Effects**

The power comparison results for homogeneous effects are summarized in Figure 3.3.

- When \( K₁ = 1, \rho_2 = 0.1 \) and \( \rho_3 = 0.1 \), the signals are super sparse and the correlation matrix has exchangeable structure. \( PC₁ \) has largest eigenvalue 4.9 and its principal angle is \( 81° \), so \( PC₁ \) has little power. The rest 39 PCs all have the same eigenvalue 0.9 and their directions are arbitrary. We observe that \( PC₂ \) also has low power. PCMinP has good power because it picks one with the smallest principal angle from the last 39 PCs. When \( K₁ = 1, \rho_2 = 0.5 \) and \( \rho_2 = 0.7 \), we observe that \( PC₂ \) is the most powerful one because \( PC₂ \) has relatively small variance 0.64 and small principal angle \( 6.6° \). \( PC₁ \) is almost powerless because it has eigenvalue (variance) 27.96 and large principal angle \( 83.4° \). All the other PCs are powerless since they have principal angles equal to \( 90° \). PCMinP also has good power since it picks \( PC₂ \) (less powerful than \( PC₂ \)) too. When \( K₁ = 1 \), MinP also has good power because it picks the smallest phenotype p-value.

- When \( K₁ = 5, \rho₁ = 0.8, \rho₂ = 0.1 \) and \( \rho₃ = 0.1 \), the signals are moderately sparse. \( PC₁ \) has eigenvalue 5.62 and principal angle \( 47° \), so it has good power. \( PC₂ \) has eigenvalue 2.97 and principal angle \( 43° \), so \( PC₂ \) is more powerful than \( PC₁ \). As
Figure 3.3: This figure shows the type I error rates and powers of the proposed methods when the genetic effects are homogeneous.

discussed above, PCBall favors the top PCs, so PCball is also very powerful because both $PC_1$ and $PC_2$ are powerful. As we decrease $\rho_1$ to 0.2 and increase $\rho_3$ to 0.8, the signals get weaker and the noises become stronger. Then $PC_1$ has very large variance 28.26 and principal angle 87°, so it’s almost powerless. In contrast, $PC_2$ has variance 1.73 and principal angle 3°, so it is much more powerful than $PC_1$. Since $PC_1$ is almost powerless, so PCBall also has very low power. PCMnP and MinP
both have good powers because they pick the most significant PC and phenotype respectively. PCFisher and PCSS which combine all the PCs waste many degrees of freedom because the signals are sparse on the phenotype and PC scales, so they are less powerful than PCMinP and MinP.

- When $K_1 = 15, 40$, the signals are dense and the power patterns change accordingly. When $K_1 = 15$, $\rho_1 = 0.2, \rho_2 = 0.1$ and $\rho_3 = 0.8$, we have weak signals since $(K_1 - 1)\rho_1 < (K_0 - 1)\rho_3$. Then $PC_1$ has very large principal angle $83^\circ$ and largest variance 20.4, so it is almost powerless. In contrast, $PC_2$ has much smaller variance 3.57 and much smaller principal angle $6.6^\circ$, so $PC_2$ is much more powerful than $PC_1$. Actually, $PC_2$ captures almost all the signal information. So PCMinP also has good power because it picks $PC_2$. However, PCSS, PCBall and PCFisher are less powerful than PCMinP because those methods waste many degrees of freedom by combing many uninformative PCs.

- When $\rho_1 = 0.8, \rho_2 = 0.1, \rho_3 = 0.1$, we have strong signals since $(K_1 - 1)\rho_1 > (K_0 - 1)\rho_3$. Then $PC_1$ has principal angle equal to $12^\circ$ and eigenvalue 12.6, so it has good power. Although $PC_2$ has has eigenvalue 2.99 but large principal angle $78^\circ$, so it’s less powerful than $PC_1$. All the other PCs have principal angles equal to $90^\circ$ and are therefore powerless. PCBall which favors the top PCs is also powerful. PCMinP also has good power because it simply picks $PC_1$. PCSS and PCFisher combine many uninformative PCs and thus are less powerful than PCMinP. When $K_1 = 40$, $PC_1$ captures all the information and other PCs are uninformative at all, so $PC_1$ and PCBall are the most powerful ones. PCMinP also has good power since it picks $PC_1$. PCSS and PCFisher are less powerful than PCMinP since they also combine many uninformative PCs. PCWSS and PCLC that both favor the last few PCs are almost powerless in these cases.

**Heterogeneous Effects**

The power results are summarized in Figure 3.4.

- When $K_1 = 5, \rho_1 = \rho_2 = \rho_3 = 0.1$, the signals are sparse and the correlation structure
Figure 3.4: This figure shows the type I error rates and powers of the proposed methods when the genetic effects are heterogeneous.

is exchangeable. $PC_1$ has eigenvalue 4.9 and principal angle 86°, so it’s almost powerless. The rest 39 PCs all have the same eigenvalue 0.9 and their directions are arbitrary. PCWSS which favors PCs with small eigenvalues has the highest power. PCSS and PCFisher which combines all the PCs (mostly from the last 39 ones) are almost as powerful as PCWSS. PCBall which favors top PCs are less powerful. PCMinP will pick the most significant PC from the last 39 ones and MinP will pick the most
significant phenotype $p$-value, so they both have good power.

- When $K_1 = 5$, $\rho_1 = 0.8$, $\rho_2 = 0.1$, $\rho_3 = 0.2$, $PC_1$ has principal angle $86^\circ$ and eigenvalue $8.23$, so $PC_1$ is still powerless. $PC_2$ has principal angle $79^\circ$ and relatively large variance $3.77$, so it is almost powerless. The last four PCs with the smallest eigenvalues $0.2$ and smallest principal angles are the most powerful PCs. So PCWSS which favors the last few PCs is very powerful. PCLC which also favors the last few PCs has good power as well. There are $34$ redundant PCs which have principal angles $90^\circ$ with eigenvalue $0.8$, so PCFisher and PCSS are less powerful because they both waste $34$ degrees of freedom. PCBall which favors the top PCs is not powerful because the first two PCs are almost powerless. PCMinP still has good power since it picks the most significant one from the last four PCs.

- When $K_1 = 5$, $\rho_1 = 0.1$, $\rho_2 = 0.1$, $\rho_3 = 0.8$, $PC_1$ has principal angle $90^\circ$ and eigenvalue $28.27$, so it is again powerless. $PC_2$ has better power than $PC_1$ because it has smaller variance $1.33$ and smaller principal angle $78.5^\circ$. The smallest eigenvalue is $0.2$ with $34$ duplicates, and its associated eigenspace is totally uninformative because it is orthogonal to the effect size vector. Actually, the eigenspace of dimension $4$ associated with eigenvalue $0.9$ is the one in which the smallest principal angle occurs. So both PCBall and PCWSS are not powerful because the most powerful PCs are neither the top nor the last ones. PCFisher and PCSS are also less powerful than PCMinP because of the wasted $34$ degrees of freedom.

- When $K_1 = 15$ and $K_1 = 40$, the signals are dense. We observe that the power patterns for $K_1 = 15$ are similar to that for $K_1 = 5$. When $K_1 = 40$, the signals are fully dense and the correlation structure is exchangeable. We observe that PCWSS is the most powerful one regardless of the value of $\rho_1$. The power of PCBall when $\rho_1 = 0.7$ is smaller than that when $\rho_1 = 0.2$ because the power of $PC_1$ deceases in $\rho_1$. Although, PCMinP and MinP are not the most powerful ones, however they still have good power and are less sensitive to the correlation structures.
### 3.5 Global Lipids GWAS Analysis

<table>
<thead>
<tr>
<th>SNP</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PCMinP</th>
<th>PCFisher</th>
<th>PCLC</th>
<th>PCBall</th>
<th>PCSS</th>
<th>PCWSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs552101</td>
<td>4.3E-09</td>
<td>0.21</td>
<td>0.71</td>
<td>0.86</td>
<td>1.7E-08</td>
<td>8.9E-07</td>
<td>0.352</td>
<td>1.24E-09</td>
<td>2.42E-07</td>
<td>0.36</td>
</tr>
<tr>
<td>rs17819328</td>
<td>4.6E-09</td>
<td>0.20</td>
<td>0.13</td>
<td>0.02</td>
<td>1.8E-08</td>
<td>6.3E-09</td>
<td>0.004</td>
<td>8.93E-10</td>
<td>5.60E-09</td>
<td>0.01</td>
</tr>
<tr>
<td>rs6029143</td>
<td>8.5E-09</td>
<td>0.65</td>
<td>0.14</td>
<td>0.70</td>
<td>3.4E-08</td>
<td>1.0E-06</td>
<td>0.071</td>
<td>3.12E-09</td>
<td>3.26E-07</td>
<td>0.32</td>
</tr>
<tr>
<td>rs9872031</td>
<td>8.6E-09</td>
<td>0.22</td>
<td>0.10</td>
<td>0.02</td>
<td>3.4E-08</td>
<td>7.6E-09</td>
<td>0.003</td>
<td>1.80E-09</td>
<td>7.03E-09</td>
<td>0.01</td>
</tr>
<tr>
<td>rs518076</td>
<td>8.8E-09</td>
<td>0.39</td>
<td>0.34</td>
<td>0.70</td>
<td>3.5E-08</td>
<td>1.5E-06</td>
<td>0.009</td>
<td>3.55E-09</td>
<td>4.59E-07</td>
<td>0.35</td>
</tr>
<tr>
<td>rs12471982</td>
<td>9.5E-09</td>
<td>0.01</td>
<td>0.07</td>
<td>0.82</td>
<td>3.8E-08</td>
<td>2.4E-08</td>
<td>0.097</td>
<td>3.55E-10</td>
<td>1.35E-08</td>
<td>0.27</td>
</tr>
<tr>
<td>rs16989081</td>
<td>1.1E-08</td>
<td>0.69</td>
<td>0.18</td>
<td>0.70</td>
<td>4.2E-08</td>
<td>1.6E-06</td>
<td>0.857</td>
<td>4.63E-09</td>
<td>4.98E-07</td>
<td>0.33</td>
</tr>
<tr>
<td>rs7203307</td>
<td>1.4E-08</td>
<td>0.60</td>
<td>0.79</td>
<td>0.75</td>
<td>5.7E-08</td>
<td>5.3E-06</td>
<td>0.018</td>
<td>8.05E-09</td>
<td>1.27E-06</td>
<td>0.31</td>
</tr>
<tr>
<td>rs4135247</td>
<td>1.8E-08</td>
<td>0.34</td>
<td>0.21</td>
<td>0.27</td>
<td>7.1E-08</td>
<td>5.9E-07</td>
<td>0.086</td>
<td>6.33E-09</td>
<td>3.43E-07</td>
<td>0.14</td>
</tr>
<tr>
<td>rs3798180</td>
<td>1.9E-08</td>
<td>0.70</td>
<td>0.10</td>
<td>0.20</td>
<td>7.5E-08</td>
<td>5.1E-07</td>
<td>0.009</td>
<td>7.59E-09</td>
<td>2.63E-07</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Table 3.1: A subset of novel SNPs identified by PCAT methods which can be detected by PC1 but no the other PCs.

<table>
<thead>
<tr>
<th>SNP</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PCMinP</th>
<th>PCFisher</th>
<th>PCLC</th>
<th>PCBall</th>
<th>PCSS</th>
<th>PCWSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7195360</td>
<td>0.04</td>
<td>0.86</td>
<td>0.05</td>
<td>3.9E-17</td>
<td>4.4E-16</td>
<td>2.7E-15</td>
<td>8.9E-12</td>
<td>0.019</td>
<td>7.3E-16</td>
<td>5.6E-17</td>
</tr>
<tr>
<td>rs41395150</td>
<td>0.37</td>
<td>0.27</td>
<td>0.89</td>
<td>4.8E-16</td>
<td>4.4E-16</td>
<td>1.1E-13</td>
<td>4.3E-16</td>
<td>0.163</td>
<td>9.6E-15</td>
<td>5.4E-17</td>
</tr>
<tr>
<td>rs10418460</td>
<td>0.73</td>
<td>0.91</td>
<td>0.70</td>
<td>3.4E-16</td>
<td>4.4E-16</td>
<td>4.9E-13</td>
<td>1.4E-14</td>
<td>0.372</td>
<td>2.0E-14</td>
<td>4.8E-17</td>
</tr>
<tr>
<td>rs2374005</td>
<td>0.89</td>
<td>0.25</td>
<td>0.15</td>
<td>2.6E-15</td>
<td>1.8E-15</td>
<td>1.6E-13</td>
<td>2.5E-16</td>
<td>0.195</td>
<td>2.9E-14</td>
<td>3.1E-16</td>
</tr>
<tr>
<td>rs6733523</td>
<td>0.58</td>
<td>0.94</td>
<td>0.27</td>
<td>2.6E-15</td>
<td>1.8E-15</td>
<td>6.2E-13</td>
<td>8.1E-13</td>
<td>0.302</td>
<td>8.2E-14</td>
<td>3.8E-16</td>
</tr>
<tr>
<td>rs697726</td>
<td>0.09</td>
<td>0.09</td>
<td>0.43</td>
<td>3.4E-15</td>
<td>4.0E-15</td>
<td>4.1E-14</td>
<td>2.4E-11</td>
<td>0.026</td>
<td>1.5E-14</td>
<td>7.0E-16</td>
</tr>
<tr>
<td>rs3001673</td>
<td>0.81</td>
<td>0.09</td>
<td>0.48</td>
<td>7.5E-15</td>
<td>4.4E-15</td>
<td>4.0E-13</td>
<td>5.9E-14</td>
<td>0.143</td>
<td>7.6E-14</td>
<td>8.8E-16</td>
</tr>
<tr>
<td>rs4694289</td>
<td>0.89</td>
<td>0.47</td>
<td>0.19</td>
<td>6.6E-15</td>
<td>4.9E-15</td>
<td>9.1E-13</td>
<td>1.8E-15</td>
<td>0.284</td>
<td>1.4E-13</td>
<td>9.3E-16</td>
</tr>
<tr>
<td>rs7658527</td>
<td>0.65</td>
<td>0.95</td>
<td>0.99</td>
<td>5.8E-15</td>
<td>4.9E-15</td>
<td>5.7E-12</td>
<td>1.8E-13</td>
<td>0.396</td>
<td>3.7E-13</td>
<td>1.0E-15</td>
</tr>
<tr>
<td>rs17189408</td>
<td>0.88</td>
<td>0.56</td>
<td>0.42</td>
<td>8.3E-15</td>
<td>6.2E-15</td>
<td>2.7E-12</td>
<td>1.2E-14</td>
<td>0.356</td>
<td>3.2E-13</td>
<td>1.3E-15</td>
</tr>
</tbody>
</table>

Table 3.2: A subset of novel SNPs identified by PCAT methods which can be detected by PC4 but not the other PCs.

To investigate the practical usefulness of the proposed principal component association tests (PCATs), we apply our testing procedures to a global lipids GWAS data set from the Global Lipids consortium (Teslovich et al., 2010). This data set contains more than 100,000 individuals of European ancestry pooled from 46 cohorts in the United States, Europe, or Australia. All the study subjects were phenotyped for four blood lipids levels: total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG). A total of about 2.6 millions directly genotyped or imputed SNPs were tested for associations with each of the four lipids in each cohort. And fixed effect meta-analysis was conducted to combine evidence for each SNP across all the cohorts. The original univariate association analysis identified 95 SNPs that
Figure 3.5: This figure shows the overlapped SNPs identified by the methods proposed in this paper. Since PCSS identified the largest number of SNPs, so we use PCSS as the reference. The top-left panel shows that PCSS found 780 novel SNPs, and 53 of them were also identified by the second PC, and 6 of them were identified by the first PC. Note that the top two PCs have nothing in common. And there are still 179 novel SNPs identified by other methods other than PCSS, PC1 and PC2. The other three panels can be easily read by following this logic. MinP is not included in this diagram because it identified zero SNP.
are associated with one or more lipids levels (Teslovich et al., 2010).

For this GWAS data set, we don’t have access to the raw individual level genotype or phenotype data, but only summary statistics containing the univariate Z-scores for each of four lipids at each genetic variant. We apply our newly developed methods to combine evidences across these four Z-scores to obtain an overall association \( p \)-value for each SNP. At the same genome-wide significance threshold level \( 5 \times 10^{-8} \), our methods identify hundreds of novel SNPs that were missed by the original univariate analysis. Each of the PCAT methods identify different sets of novel SNPs that haven’t been reported yet. Some of the novel SNPs are detected by more than one methods. For example, SNP rs552101 is identified by \( PC_1 \), PCMinP and PCBall methods. And some SNPs can only be detected by one method, for example, SNP rs501470 is only detected by PCBall. This shows that our PCAT methods can greatly increase statistical powers to detect genetic variants by leveraging the correlations among the four lipids.

We also observe that each of our PCAT has their own merit in certain situations and no single method is uniformly better than the others. For this particular data set, PCSS, PCFisher and PCMinP methods identify 780, 747 and 375 novel SNPs respectively. For comparison purpose, we also include MinP method which doesn’t detect any genome-wide significant hit. This is because MinP can only pick the smallest univariate \( p \)-value and adjust for multiple testing accordingly. It will lose substantial power if none of the four \( p \)-values is significant. On the contrary, PCAT methods can overcome this drawback and detect the associations on the PC scale by orthogonal transformations of the original Z-scores. Association signals are too weak to be detected on the original scale could be well detected on the PC scale. This is the underlying reason why the principal component association tests are more powerful than MinP and univariate analysis approaches.

To compare the PCAT methods, we take two subsets of novel SNPs identified by our methods and present them in Table 3.1 and 3.2. None of the four Z-scores for these novel SNPs reached genome-wide significance threshold, so these SNPs can’t be detected by any univariate analysis methods. We now show that different PCAT methods can capture different signals. First, the four estimated eigenvalues are \( \lambda = (2.09, 1.32, 0.54, 0.05) \) and
the four estimated eigenvectors are:

\[ \mathbf{u}_1^T = (0.12, -0.63, -0.64, -0.43), \quad \mathbf{u}_2^T = (0.78, 0.19, 0.30, -0.51), \]
\[ \mathbf{u}_3^T = (-0.55, 0.44, -0.06, -0.71), \quad \mathbf{u}_4^T = (0.27, 0.61, -0.71, 0.23). \]

Table 3.1 displays ten SNPs associated with \( PC_1 \), but not with the other three PCs. The reason is that the Z-scores have small principal angles with the first eigenvector. Consequently, other combination methods that favor \( PC_1 \) are also more likely to be significant, such as PCBall. However, PCWSS and PCLC which favor the last PCs are not significant in this case. For example, SNP rs552101 has Z-scores \( (2.268, -5.228, -4.931, -4.193)^T \). If we apply \( PC_1 \) to detect this SNP, then \( PC_1 = \mathbf{u}_1^T \mathbf{Z} = 8.525 \) with variance 2.09, so \( PC_1 \) has \( p \)-value equal to \( 4.3 \times 10^{-9} \). However, if we apply \( PC_4 \), then \( PC_4 = \mathbf{u}_4^T \mathbf{Z} = -0.04 \) with variance 0.05, so \( PC_4 \) has \( p \)-value 0.86. As shown above, whenever \( PC_1 \) is significant, PCBall which favors \( PC_1 \) also tends to be significant. But PCWSS which favors the last PC is not significant. PCMinP which will pick \( PC_1 \) also has significant \( p \)-value. As for PCFisher, PCLC and PCSS which all favor PCs with small eigenvalues tend to be less or not significant.

Table 3.2 displays another set of ten SNPs which can be detected by \( PC_4 \) but not the other three PCs. There are two reasons: (1) the Z-scores for those SNPs have smaller empirical angles with the fourth eigenvector (about 40° compared to about 70° of the first three eigenvectors); (2) the eigenvalue is 0.05 which is much smaller than the first three eigenvalues. For example, SNP rs7195360 has Z-scores \( Z = (0.75, -3.632, -0.444, -0.8)^T \), and the fourth eigenvector is \( \mathbf{u}_4 = (0.27, 0.61, -0.71, 0.23)^T \). Thus, \( PC_4 = \mathbf{u}_4^T \mathbf{Z} = -1.88 \) with variance 0.05, and its \( p \)-value is \( 3.9 \times 10^{-17} \). Since \( PC_4 \) is highly significant, therefore PCWSS is also significant, as well as PCFisher, PCLC and PCSSS. But PCBall which doesn’t favor \( PC_4 \) is not significant. PCMinP simply picks \( PC_4 \) and is also significant.

### 3.6 Discussion

In this paper, we propose several principal component association testing (PCAT) procedures to detect pleiotropic effects in multiple phenotype association studies. Our methods
leverage the correlation structures among the multiple phenotypes to increase statistical powers. We show both analytically and through simulations when the PCAT methods will be powerful and when they are not. The results can help better guide researchers to choose proper tests for their own studies. We re-analyzed a global lipids GWAS data set and identified hundreds of novel SNPs missed by previous univariate analysis, which demonstrates the practical potentials of the proposed methods.

PCAT methods are applicable to both individual level data and summary statistics, and therefore provide great flexibility for various practical situations. Moreover, PCAT methods can coherently integrate both continuous and binary phenotypes to increase power without modeling their joint distribution explicitly. This is especially useful for complex diseases and traits in many biomedical research where both binary and quantitative traits are collected. By introducing the concept of principal angle, we are able to conduct theoretical eigen-analysis and power analysis from a geometric perspective. We also explain how the correlation structures, dimension of the phenotypes, signal sparsity and effect homogeneity all together influence the powers of PCAT methods. Besides, we provide analytic formulas to compute the $p$-values for all of the PCAT methods without resorting to computational intensive methods like permutations or simulations. This is especially attractive in large-scale genetic association studies where millions of genetic variants are screened with tens or even hundreds of phenotypes and therefore computational efficiency is of top priority. We further conduct simulation studies to demonstrate that our methods maintain correct type I error rates and their powers are compared under a wide range of settings. An application to a global lipids GWAS data set illustrates the practical usefulness of our methods for multiple phenotype association studies.

In practice, researchers usually have no prior information about the underlying relationship between a genetic variant and multiple phenotypes. So it might be difficult to choose the best method for a specific data set. We hereby recommend more robust methods such as PCFisher for the first round association screening. Although we use principal component axis of the summary statistics as the reference directions for the orthogonal transformation, however, this axis system might not be the best one to capture the signals since the effect size vector might not align with any of the PC axis. Therefore, future research is
needed to improve current PC based methods. For example, instead of using the unsupervised PC axis system, it might be better to find a supervised orthogonal transformation that takes the effect size vector into account explicitly.
References


that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metabolism* 9(4), 311–326.


Appendix A

Proof of Independence of Two Score Statistics

A.1 Part a

Using the following result from the properties of multivariate normal distributions which states that if $Z$ is a multivariate normal random vector with covariance matrix $\Sigma$, then $AZ$ and $BZ$ are independent of each other if and only if $\text{Cov}(AZ, BZ) = A\Sigma B^T = 0$.

In this case, $A = J^T\Sigma^{-1}$, and $B = \Sigma^{-1}(I - H)$, where $H = J(J^TJ)^{-1}J^T$. So we have

\[
\text{Cov}(AZ, BZ) = A\Sigma B^T = J^T\Sigma^{-1}\Sigma(\Sigma^{-1}(I - H))^T
\]

\[
= J^T\Sigma^{-1}\Sigma(I - H)\Sigma^{-1}
\]

\[
= J^T(I - J(J^TJ)^{-1}J^T)\Sigma^{-1}
\]

\[
= (J^T - J^T)\Sigma^{-1}
\]

\[
= 0.
\]

Therefore, $U_{\mu_0}$ and $\Sigma^{-1}(I - H)Z$ are independent of each other, and note that $U_{\mu_0}$ is independent of any measurable function of the elements of $\Sigma^{-1}(I - H)Z$. Hence, this proves that $U_{\mu_0}$ and $U_{\tau_0}$ are statistically independent.
A.2 Part b

We could also prove that $U_{\mu_0}^2$ and $U_{\tau_0}$ are independent directly using results from quadratic form of norm random variables.

The proof is based on the Craig’s theorem (Craig, 1943) which says that $X^TAX$ and $X^TBX$ are independent if and only if $A\Sigma B = 0$, given that $X \sim N(\mu, \Sigma)$. Therefore we only need to check whether $\Lambda_{\mu_0} \Sigma \Lambda_{\tau_0} = 0$, where $\Lambda_{\mu_0} = \Sigma^{-1}J J^T \Sigma^{-1}$ and $\Lambda_{\tau_0} = (I - H) \Sigma^{-1} \Sigma^{-1}(I - H)$. So we have

\[
\Lambda_{\mu_0} \Sigma \Lambda_{\tau_0} = \Sigma^{-1}11^T \Sigma^{-1} \Sigma (I - H) \Sigma^{-1} \Sigma^{-1}(I - H)
\]

\[
= \Sigma^{-1}11^T (I - H) \Sigma^{-1} \Sigma^{-1}(I - H)
\]

\[
= \Sigma^{-1}1 (I^T - I^T) \Sigma^{-1} \Sigma^{-1}(I - H)
\]

\[
= 0.
\]

Therefore $U_{\mu_0}^2$ and $U_{\tau_0}$ are independent.
Appendix B

Derivation of the Eigenvalues of Block Exchangeable Correlation Matrices

B.1 Eigen-analysis

When \( \rho_2 \neq 0 \), the eigenvalues of the matrix \( \Sigma \) are the roots of the following characteristic equation

\[
\det (\Sigma - \lambda I) = 0. \tag{B.1}
\]

And we can represent \( \Sigma - \lambda I \) as a block matrix given by

\[
\Sigma - \lambda I = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}, \tag{B.2}
\]

where

\[
A = \Sigma_1 - \lambda I_{K_1}, \quad B = \Sigma_2, \quad C = \Sigma_3 - \lambda I_{K_0}.
\]

If \( \lambda \neq 1 - \rho_1 \) and \( \lambda \neq 1 + (K_1 - 1)\rho_1 \), then \( A \) is non-singular. Therefore we can invert \( A \) and have the following identity

\[
\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ B^T & I \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & D - B^T A^{-1}B \end{pmatrix}. \tag{B.3}
\]

Therefore we obtain that

\[
\det \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} = \det(A) \det(C - B^T A^{-1}B). \tag{B.4}
\]
Since $A$ has exchangeable structure, we can easily derive its determinant and its inverse matrix $A^{-1}$ using Sylvester’s determinant theorem (Akritas et al., 1996)

$$\det(\Sigma_{11} - \lambda I_{K_1}) = \det(A) = (1 - \lambda - \rho_1)^{K_1 - 1} (1 - \lambda + (K_1 - 1)\rho_1). \quad (B.5)$$

To calculate the inverse matrix of $A$, we employ Sherman-Morrison formula (Sherman and Morrison, 1949, 1950) to obtain that

$$A^{-1} = \left( (1 - \rho_1 - \lambda)I_{K_1} + \rho_1 1_{K_1} 1_{K_1}^T \right)^{-1} \left(1 - \rho_1 - \lambda\right)^{K_1 - 1} \left(1 - \rho_1 - \lambda + (K_1 - 1)\rho_1\right)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1)\right)^{K_1 - 1} (1 - \lambda + (K_1 - 1)\rho_1)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1)\right)^{K_1 - 1} (1 - \lambda + (K_1 - 1)\rho_1)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1). \quad (B.6)$$

where $a = \frac{1}{1 - \rho_1 - \lambda}$ and $b = \frac{\rho_1}{(1 + \frac{K_1 - 1}{1 - \rho_1 - \lambda}) (1 - \lambda + (K_1 - 1)\rho_1)^2}$. After substituting $A^{-1}$ back into $C - B^T A^{-1} B$, we have

$$C - B^T A^{-1} B = C - \rho_2^2 1_{K_0} 1_{K_1}^T A^{-1} 1_{K_0} 1_{K_1}^T.$$

$$= C - \rho_2^2 1_{K_0} \left( a K_1 - b K_1^2 \right) 1_{K_0}^T.$$

$$= (1 - \rho_3 - \lambda) I_{K_0} + (\rho_3 - \rho_2^2 a K_1 + \rho_2^2 b K_1^2) 1_{K_0} 1_{K_0}^T.$$

$$= (1 - \rho_3 - \lambda) I_{K_0} + d 1_{K_0} 1_{K_0}^T. \quad (B.7)$$

where $d = \rho_3 - \rho_2^2 a K_1 + \rho_2^2 b K_1^2$. Therefore using Sylvester’s determinant theorem again (Akritas et al., 1996), we have

$$\det(C - B^T A^{-1} B) = (1 - \rho_3 - \lambda)^{K_0 - 1} (1 - \rho_3 - \lambda + K_0 d). \quad (B.8)$$

Combining these results together, we have

$$\det(\Sigma - \lambda I)$$

$$= \left(1 - \lambda - \rho_1\right)^{K_1 - 1} \left(1 - \lambda + (K_1 - 1)\rho_1\right)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1)\right)^{K_1 - 1} \left(1 - \lambda + (K_1 - 1)\rho_1\right)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1)\right)^{K_1 - 1} \left(1 - \lambda + (K_1 - 1)\rho_1\right)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1)\right)^{K_1 - 1} \left(1 - \lambda + (K_1 - 1)\rho_1\right)^{K_0 - 1} (1 - \lambda + (K_1 - 1)\rho_1). \quad (B.9)$$

where $d$ depends on $\lambda$, and $l = 1 - \lambda + (K_1 - 1)\rho_1$. From equation (B.9), we know one eigenvalue is $1 - \rho_1$ with algebraic multiplicity $K_1 - 1$ and another eigenvalue is $1 - \rho_3$ with
algebraic multiplicity $K_0 - 1$. The third factor in (B.9) is a quadratic function in $\lambda$ which can be easily solved for its roots. To be more concrete, the third factor is represented as
\[
[(1 - \rho_3 - \lambda)l + K_0 \rho_3 l - \rho_2^2 K_1 K_0] = \lambda^2 + g\lambda + h,
\]
where $g$ and $h$ are given by
\[
g = -[2 + (K_1 - 1)\rho_1 + (K_0 - 1)\rho_3]
\]
\[
h = 1 - \rho_3 + (K_1 - 1)\rho_1(1 - \rho_3) + K_0 \rho_3 + K_0 \rho_3(K_1 - 1)\rho_1 - \rho_2^2 K_1 K_0
\]
Since the discriminant of a real symmetric matrix is always non-negative, so we require that $g < 0$ and $h > 0$ to ensure that the roots of this quadratic equation are all positive. Some violations could happen, for instance, if we specify $K_1 = 4$, $K_0 = 3$, $\rho_1 = 0.5$, $\rho_2 = 0.9$ and $\rho_3 = 0.6$, then we obtain a negative eigenvalue $-0.77$, which indicates that the specified matrix is not a valid correlation matrix. So the two positive eigenvalues that solve this quadratic equation are given by
\[
\lambda_j = \frac{1}{2} (-g \pm \sqrt{g^2 - 4h}), \quad j = 1, 2,
\]
each of them has algebraic multiplicity 1. Now we summarize the eigenvalue formula as
\[
\lambda_k = \begin{cases} 
\frac{1}{2} (-g + \sqrt{g^2 - 4h}) & k = 1 \\
\frac{1}{2} (-g - \sqrt{g^2 - 4h}) & k = 2 \\
1 - \rho_1 & 3 \leq k \leq K_1 + 1 \\
1 - \rho_3 & K_1 + 2 \leq k \leq K
\end{cases}
\]
Note that in (B.11), those eigenvalues are not sorted.