

2013

Statistical methods for genetic association studies: multi-cohort and rare genetic variants approaches

<https://hdl.handle.net/2144/13141>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**STATISTICAL METHODS FOR GENETIC ASSOCIATION STUDIES:
MULTI-COHORT AND RARE GENETIC VARIANTS APPROACHES**

by

HAN CHEN

B.S., Tsinghua University, 2007
M.A., Columbia University, 2009

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2013

© 2013 by
HAN CHEN
All rights reserved

Approved by

First Reader

Josée Dupuis, Ph.D.
Professor of Biostatistics

Second Reader

Ching-Ti Liu, Ph.D.
Assistant Professor of Biostatistics

Third Reader

Qiong Yang, Ph.D.
Associate Professor of Biostatistics

To my beloved late grandparents

Yaxun Zhu (1926 – 2001)

Xingyi Zheng (1925 – 2009)

Yimin Chen (1923 – 2013)

Acknowledgments

I would like to thank all my teachers on my long journey towards this degree. My deepest gratitude goes to my major advisor Dr. Josée Dupuis for her encouragement and guidance in my dissertation research. This dissertation could not be finished without her expertise in statistics and genetics, and her insightful thoughts.

I am very grateful to my academic advisor Dr. Ching-Ti Liu for his generous help and constant support in my graduate study and research. I would also like to thank Dr. Qiong Yang, Dr. Michael LaValley and Dr. Serkalem Demissie for sharing their ideas and offering advice in my dissertation research.

Finally, I would like to thank my parents Yueqi Chen and Deshu Xu, and my wife Mengxi Wang, for their unconditional love, support and encouragement.

real data examples. First, we develop a method of moments estimator for the between-study covariance matrix in random effects model multivariate meta-analysis. Our estimator is the first such estimator in matrix form, and holds the invariance property to linear transformations. It has similar performance with existing methods in simulation studies and real data analysis. Next, we extend the Sequence Kernel Association Test (SKAT), a rare genetic variants analysis approach for unrelated individuals, to be applicable in family samples for quantitative traits. The extension is necessary, as the original test has inflated type I error when directly applied to related individuals, and selecting an unrelated subset from family samples reduces the sample size and power. Finally, we derive methods for rare genetic variants analysis in detecting gene by environment interaction on quantitative traits, in the context of univariate test on the interaction term parameter. We develop statistical tests in the settings of both burden test and SKAT, for both unrelated and related individuals. Our methods are relevant to genetic association studies, and we hope that they can facilitate research in this field and beyond.

Table of Contents

Chapter 1 Introduction	1
1.1 Genetic Association Studies.....	1
1.2 Genetic Association Studies in Family Samples.....	3
1.3 Meta-Analysis	3
1.4 Rare Genetic Variants Analysis	7
1.5 Gene by Environment Interaction	9
1.6 Dissertation Outline.....	10
Chapter 2 A Method of Moments Estimator in Random Effects Multivariate Meta-Analysis	12
2.1 Introduction	12
2.2 Fixed Effect Multivariate Meta-Analysis.....	14
2.3 Random Effect Multivariate Meta-Analysis	16
2.3.1 Restricted Maximum Likelihood Method	17
2.3.2 Jackson’s Multivariate DerSimonian and Laird’s Method.....	18
2.3.3 Chen’s Multivariate DerSimonian and Laird’s Method.....	18
2.3.3.1 Invariance Property to Linear Transformations	19
2.3.3.2 Connection with Univariate DerSimonian and Laird’s Estimator	21
2.3.3.3 Homogeneity Test in the Presence of Heterogeneity.....	22
2.3.4 Positive Semi-Definiteness of Covariance Matrix Estimators	23
2.4 Simulation Studies.....	24
2.4.1 Simulation Design	24

2.4.2 Simulation Results	26
2.4.3 Additional Simulation Studies.....	36
2.5 Application	38
2.6 Discussion	41
Chapter 3 Sequence Kernel Association Test for Quantitative Traits in Family	
Samples	43
3.1 Introduction	43
3.2 Burden Test for Quantitative Traits in Family Samples	45
3.3 Sequence Kernel Association Test for Quantitative Traits in Family Samples	46
3.3.1 Connection with SKAT in Unrelated Individuals	48
3.3.2 Reparametrization.....	49
3.4 Simulation Studies.....	50
3.4.1 Type I Error	50
3.4.1.1 Simulation Design.....	50
3.4.1.2 Simulation Results	51
3.4.2 Power	55
3.4.2.1 Simulation Design.....	55
3.4.2.2 Simulation Results	56
3.5 Analysis of Framingham Heart Study Data	58
3.5.1 Candidate Gene Study	58
3.5.2 Sliding Window Analysis.....	62
3.6 Computation Time.....	64

3.7 Discussion	65
Chapter 4 Methods for Rare Genetic Variants Analysis in Detecting Gene by Environment Interaction on Quantitative Traits.....	70
4.1 Introduction	70
4.2 Burden Test for Gene-Environment Interaction.....	72
4.3 Sequence Kernel Association Test for Gene-Environment Interaction	73
4.3.1 Fixed Main Effects	73
4.3.2 Random Main Effects with Residuals Adjusting for Covariates Only.....	74
4.3.3 Random Main Effects with Residuals Adjusting for Genotype Main Effects .	75
4.4 Extension to Related Individuals.....	76
4.4.1 Burden Test.....	76
4.4.2 SKAT with Fixed Main Effects.....	77
4.4.3 SKAT with Random Main Effects	78
4.5 Simulation Studies.....	80
4.5.1 Type I Error	80
4.5.1.1 Simulation Design.....	80
4.5.1.2 Simulation Results	82
4.5.2 Power	87
4.5.2.1 Simulation Design.....	87
4.5.2.2 Simulation Results	90
4.6 Application	93
4.7 Discussion	95

Chapter 5 Summary and Future Work	98
5.1 Summary	98
5.2 Future Work	99
5.2.1 Extension of the Method of Moments Estimator.....	99
5.2.2 Sequence Kernel Association Test for Dichotomous Traits in Family Samples	100
5.2.3 Joint Test of Genetic Main Effects and Gene by Environment Interaction for Rare Genetic Variants.....	101
Appendices.....	103
Appendix A Derivation of the Method of Moments Estimator	103
Bibliography	108
Curriculum Vitae	117

List of Tables

Table 2.1	Simulation results for summary effect estimates (between-study correlation 0.2, within-study correlation 0.2).....	27
Table 2.2	Simulation results for between-study covariance matrix (between-study correlation 0.2, within-study correlation 0.2)	28
Table 2.3	Simulation results for summary effect estimates (between-study correlation 0.2, within-study correlation 0.8).....	30
Table 2.4	Simulation results for between-study covariance matrix (between-study correlation 0.2, within-study correlation 0.8)	31
Table 2.5	Simulation results for summary effect estimates (between-study correlation 0.8, within-study correlation 0.2).....	32
Table 2.6	Simulation results for between-study covariance matrix (between-study correlation 0.8, within-study correlation 0.2)	33
Table 2.7	Simulation results for summary effect estimates (between-study correlation 0.8, within-study correlation 0.8).....	34
Table 2.8	Simulation results for between-study covariance matrix (between-study correlation 0.8, within-study correlation 0.8)	35
Table 2.9	Simulation results from various numbers of studies (proportions of marginal variation due to heterogeneity 0.5, between-study correlation 0.2, within-study correlation 0.2).....	38
Table 2.10	Regression results from 8 race groups	40
Table 2.11	Meta-analysis results for 8 race groups	41

Table 3.1	Type I errors of famSKAT, famBT, unrSKAT and SKAT	54
Table 3.2	Candidate gene study results from unrSKAT	60
Table 3.3	Candidate gene study results from famSKAT and famBT	61
Table 3.4	Comparison of Kuonen's and Davies' methods in calculating p-values in the tail	69
Table 4.1	Type I errors from the null simulation without genotype main effects	83
Table 4.2	Type I errors from the null simulation with genotype main effects	84

List of Figures

Figure 2.1 Simulation results from various numbers of studies (proportions of marginal variation due to heterogeneity 0.5, between-study correlation 0.2, within-study correlation 0.2).....	37
Figure 3.1 Distribution of null p-values of famSKAT, famBT, unrSKAT and SKAT ..	53
Figure 3.2 Power comparisons of famSKAT, famBT and unrSKAT.....	57
Figure 3.3 Quantile-Quantile plots for famSKAT in the genome-wide sliding window analysis on four glycemic traits	63
Figure 3.4 Run time of famSKAT, famBT and SKAT in analyzing 20 SNPs	65
Figure 4.1 Quantile-Quantile plots from the null simulation without genotype main effects	85
Figure 4.2 Quantile-Quantile plots from the null simulation with genotype main effects in the same direction	86
Figure 4.3 Quantile-Quantile plots from the null simulation with genotype main effects in different directions	87
Figure 4.4 Power comparisons of SKAT-FIX, SKAT-RAN and BT in detecting gene by BMI interaction.....	92
Figure 4.5 Power comparisons of SKAT-FIX, SKAT-RAN and BT in detecting gene by sex interaction	93
Figure 4.6 Quantile-Quantile plots for SKAT-type tests in the genome-wide sliding window analysis for gene by BMI interaction on fasting glucose.....	94

List of Abbreviations

BMI	Body Mass Index
BT	Burden Test
Chr	Chromosome
CMC	Combined Multivariate and Collapsing
CPU	Central Processing Unit
EM	Expectation-Maximization
famBT	Family-Sample Burden Test
famSKAT	Family-Sample Sequence Kernel Association Test
FEMA	Fixed Effects Meta-Analysis
FHS	Framingham Heart Study
GC	Genomic Control
GLS	Generalized Least Squares
GWAS	Genome-Wide Association Studies
HOMA-B	Homeostatic Model Assessment for β -Cell Function
HOMA-IR	Homeostatic Model Assessment for Insulin Resistance
HSLs	High School Longitudinal Study
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MDLC	Multivariate DerSimonian and Laird's Method by Chen et al.
MDLJ	Multivariate DerSimonian and Laird's Method by Jackson et al.
MSE	Mean Squared Error

OLS	Ordinary Least Squares
REML	Restricted Maximum Likelihood
SE	Standard Error
SHARe	Single Nucleotide Polymorphism Health Association Resource
SKAT	Sequence Kernel Association Test
SKAT-FIX	Sequence Kernel Association Test with Fixed Main Effects
SKAT-O	Sequence Kernel Association Test – Optimal Test
SKAT-RAN	Sequence Kernel Association Test with Random Main Effects
SNP	Single Nucleotide Polymorphism
unrSKAT	Unrelated-Sample Sequence Kernel Association Test

Chapter 1 Introduction

1.1 Genetic Association Studies

Genetic association studies, including genome-wide association studies (GWAS) and candidate gene studies, have been widely used to identify genetic markers associated with complex diseases or disease-related quantitative traits. In GWAS, investigators are usually interested in biallelic common genetic variants, defined as genetic markers with minor allele frequency (MAF) greater than 5% or 1%, using a single marker test such as logistic regression or linear regression. In recent years, a large number of genetic markers associated with complex diseases and related quantitative traits have been identified by GWAS. However, they only explain a small proportion of heritability in these traits, and different strategies for exploring the unexplained heritability are of great interest in this field [Eichler et al., 2010].

One reason is that the sample size of a single cohort or case-control study is not large enough to detect association with genetic variants with smaller effect sizes. As the sample size increases, genetic variants with weaker association can be found due to increased power [McCarthy et al., 2008]. Thus, consortia have been formed to unite multiple cohorts or case-control studies [Zeggini et al., 2008; Psaty et al., 2009]. Investigators conduct association analysis separately and combine their study results via meta-analysis, and they have identified a lot of novel genetic variants associated with different traits [Dupuis et al., 2010; Lindgren et al., 2009; Prokopenko et al., 2009]. Lin and Zeng [2010]

showed that meta-analysis does not have compromised power compared with combining individual data. Thus, collaborating with other cohorts to increase the sample size by meta-analysis is one possible way to explore the unexplained heritability.

On the other hand, the single marker test in GWAS is not powerful in detecting the association with rare genetic variants. However, there are a lot more rare genetic variants than common genetic variants in the human genome, and they may explain some of the unexplained heritability. With the advances in sequencing technology and decreasing cost, rare genetic variants data have become available in many cohorts. As an emerging field in genetic association studies, rare genetic variants analysis has become of great interest in recent years.

Genetic markers are not the only determinant for complex diseases or disease-related quantitative traits. Epidemiological studies have identified the association between diseases and environmental variables. Thus, it is reasonable to hypothesize that environmental variables may interact with genetic markers on the traits. Statistical approaches for detecting gene by environment interaction have been proposed for single cohort [Kraft et al., 2007], and also in the meta-analysis context [Manning et al., 2011]. They have been widely used to identify novel association and to help explore the unexplained heritability [Manning et al., 2012].

1.2 Genetic Association Studies in Family Samples

Family-based study designs have a long history in linkage analysis of diseases and quantitative traits [Falk and Rubinstein, 1987; Ott, 1989; Terwilliger and Ott, 1992; Spielman, McGinnis and Ewens, 1993]. However, in GWAS era, researchers have become more interested in unrelated individuals in genetic association studies. For family samples, ordinary regression approaches are not directly applicable, because inflated type I error is observed when familial correlation is not appropriately modeled. A simple way to solve this issue, selecting unrelated individuals from family samples, usually leads to great power loss due to reduced sample size.

Statistical approaches have been proposed to analyze correlated data. In the GWAS context, for quantitative traits, linear mixed effects models that take familial correlation as a random effect with covariance proportional to the kinship matrix is commonly used for single marker tests [Amos, 1994; Almasy and Blangero, 1998; Pankratz, de Andrade and Therneau, 2005]. For dichotomous traits, generalized estimating equations [Liang and Zeger, 1986] are usually applied in genetic association studies to analyze family data.

1.3 Meta-Analysis

Meta-analysis is a general approach to combine evidence from multiple studies to make inference about one or more effect sizes of interest. Meta-analysis methods include p-value based approaches and effect size based approaches. For p-value based approaches,

the p-values from all k studies p_i ($1 \leq i \leq k$) are combined into an overall p-value.

Fisher [Fisher, 1925] proposed using the statistic

$$T_1 = \sum_{i=1}^k -2 \log p_i .$$

Under the null hypothesis, all p_i should follow a uniform distribution on the interval (0, 1), and are independent provided that the samples do not overlap between studies. Thus T_1 follows a chi-square distribution with $2k$ degrees of freedom. Alternatively, we can use Stouffer's method [Stouffer et al., 1949], resulting in the test statistic

$$T_2 = \frac{1}{\sqrt{k}} \sum_{i=1}^k \Phi^{-1}(p_i) ,$$

where Φ is the cumulative distribution function of standard normal distribution. Under the null hypothesis, T_2 follows a standard normal distribution.

The major drawback of p-value based approaches is that they do not take the direction of association into consideration. A weighted signed Z-score approach [Lipták, 1958] converts the p-value p_i from each study to a standard normal statistic Z_i with the sign reflecting the direction of association, and calculates the test statistic

$$T_3 = \frac{\sum_{i=1}^k \sqrt{n_i} Z_i}{\sqrt{\sum_{i=1}^k n_i}} ,$$

where n_i is the sample size in study i . Under the null hypothesis, T_3 also follows a standard normal distribution. However, we cannot get a summary effect size estimate using the three methods discussed above.

Depending on the number of effect sizes of interest, effect size based approaches can be categorized into univariate meta-analysis and multivariate meta-analysis. In both cases, inverse variance (or covariance matrix) weighted meta-analysis using fixed effects model and random effects model have been proposed. The validity of the fixed effects model meta-analysis depends on the underlying assumption that all studies in the meta-analysis share the same effect size. In the presence of heterogeneity, the fixed effects model incorrectly ignores the between-study variance (or covariance matrix) and may yield false positive results. The random effects model takes into account both within-study and between-study variances (or covariance matrices). It is more conservative than the fixed effects model and should be favored in the presence of heterogeneity.

In the univariate case, we use the effect estimate b_i and its standard error $SE(b_i)$ from each study and calculate the fixed effects model summary effect estimate

$$\hat{\beta}_F = \frac{\sum_{i=1}^k \frac{b_i}{SE(b_i)^2}}{\sum_{i=1}^k \frac{1}{SE(b_i)^2}},$$

with standard error

$$SE(\hat{\beta}_F) = \frac{1}{\sqrt{\sum_{i=1}^k \frac{1}{SE(b_i)^2}}}.$$

We can also perform the random effects model meta-analysis by calculating the summary effect estimate

$$\hat{\beta}_R = \frac{\sum_{i=1}^k \frac{b_i}{SE(b_i)^2 + \hat{\tau}^2}}{\sum_{i=1}^k \frac{1}{SE(b_i)^2 + \hat{\tau}^2}},$$

with standard error

$$SE(\hat{\beta}_R) = \frac{1}{\sqrt{\sum_{i=1}^k \frac{1}{SE(b_i)^2 + \hat{\tau}^2}}},$$

where the between-study variance estimate $\hat{\tau}^2$ is estimated using the restricted maximum likelihood (REML) method, the expectation-maximization (EM) algorithm [Dempster, Laird and Rubin, 1977], DerSimonian and Laird's [1986] method of moments, or other approaches.

In the context of multivariate meta-analysis, we use the effect estimate \mathbf{b}_i and its covariance matrix estimate $Cov(\mathbf{b}_i)$ from each study and calculate the fixed effects model summary effect estimate

$$\hat{\beta}_F = \left(\sum_{i=1}^k Cov(\mathbf{b}_i)^{-1} \right)^{-1} \left(\sum_{i=1}^k Cov(\mathbf{b}_i)^{-1} \mathbf{b}_i \right),$$

with covariance matrix estimate

$$Cov(\hat{\beta}_F) = \left(\sum_{i=1}^k Cov(\mathbf{b}_i)^{-1} \right)^{-1}.$$

If we perform a random effects model multivariate meta-analysis, we first get an estimate for the between-study covariance matrix $\hat{\mathbf{T}}$, then we can calculate the summary effect estimate

$$\hat{\boldsymbol{\beta}}_R = \left(\sum_{i=1}^k (\text{Cov}(\mathbf{b}_i) + \hat{\mathbf{T}})^{-1} \right)^{-1} \left(\sum_{i=1}^k (\text{Cov}(\mathbf{b}_i) + \hat{\mathbf{T}})^{-1} \mathbf{b}_i \right),$$

with covariance matrix estimate

$$\text{Cov}(\hat{\boldsymbol{\beta}}_R) = \left(\sum_{i=1}^k (\text{Cov}(\mathbf{b}_i) + \hat{\mathbf{T}})^{-1} \right)^{-1}.$$

The fixed effects model is more widely used than the random effects model in multivariate meta-analysis, possibly due to the difficulty in estimating the between-study covariance matrix.

1.4 Rare Genetic Variants Analysis

When analyzing rare genetic variants, the single marker test, which is typically used for common genetic variants analysis in GWAS, is not powerful. Supposing that we have q rare genetic variants in a genomic region, the multivariate test in a regression framework is also not ideal because of the large number (q) of parameters. To jointly analyze multiple rare genetic variants and to reduce the number of parameters, various burden tests have been proposed [Li and Leal, 2008; Morgenthaler and Thilly, 2007; Madsen and Browning, 2009; Morris and Zeggini, 2010]. These tests first calculate the combined genetic score for all rare genetic variants, and then perform a univariate test on the combined genotype score. We can simply use an indicator of any rare variants, calculate the total number (or proportion) of rare alleles, or use a weighted sum. Burden tests are most powerful when all rare genetic variants in the test share the same direction of effects, or even have the same effect size. If there are both protective and detrimental rare genetic

variants in the test, the effects would cancel out, which significantly decreases the power of burden tests.

Li and Leal [2008] proposed a compromise between burden tests and multivariate tests. Their approach, named combined multivariate and collapsing (CMC) method, divides the genetic variants analyzed into several categories based on the MAF spectra and assesses association between the trait and multiple combined genotype scores. The CMC method performs better than burden tests in certain scenarios because it allows different genotype scores to have different directions of effects. However, it works best when all genetic variants in each category share the same direction of effects, which may not be true in practice.

Han and Pan [2010] proposed a data-adaptive sum test which does not make any assumption on the direction of effects. Instead, it first performs single marker tests on all rare genetic variants analyzed and then calculates the combined genotype score using the signs from single marker tests. Hoffman, Marini and Witte [2010] developed a step-up approach which not only takes the signs, but also incorporates weights. However, since both methods use data-driven information in the test, they require permutation to calculate p-values.

Wu et al. [2011] proposed the sequence kernel association test (SKAT) as a non-collapsing method in rare genetic variants analysis. It does not make any assumptions on

the direction of effects. It treats genotype effects as random effects with mean 0 and tests the variance component, which greatly reduces the number of parameters in the test, compared with multivariate tests. Assuming the individuals are unrelated, SKAT is a flexible and computationally efficient approach. Moreover, it calculates the p-value analytically without permutation.

1.5 Gene by Environment Interaction

Gene by environment interaction has become of great interest in genetic association studies. Kraft et al. [2007] proposed general approaches for gene by environment interaction analysis in the case-control context. Moreover, their approach can also be applied for quantitative traits. Without loss of generality, we assume a single marker test with a common genetic variant G and only one environmental covariate E . We are interested in testing the interaction of G and E on the trait y :

$$g(E(y)) = \beta_0 + \beta_1 E + \beta_2 G + \beta_3 EG .$$

One strategy is to test a single parameter $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$, using either Wald, likelihood ratio, or score test. This is an interaction-only test. Another approach is to perform a joint test on the genetic main effect and gene by environment interaction, corresponding to $H_0: \beta_2 = \beta_3 = 0$ versus $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$. By performing the interaction-only test we are interested in the gene by environment interaction, and by performing the joint test we are actually testing if there is any genetic association with the marker, allowing for gene by environment interaction.

Gene by environment interaction analysis has also been proposed in the meta-analysis context [Manning et al., 2011]. Since the effect size of genetic markers are often of great interest in genetic association studies, effect size based meta-analysis approaches are preferred over p-value based approaches. When we perform an interaction-only test, we would use univariate meta-analysis; when we perform a joint test on the genetic main effect and gene by environment interaction, we would use bivariate meta-analysis. In both cases, the fixed effects model is widely applied, albeit it is not valid in the presence of heterogeneity.

1.6 Dissertation Outline

In this dissertation, we develop novel statistical methods to facilitate research in genetic association studies. Our work involves meta-analysis, family samples, rare genetic variants analysis, and gene by environment interaction analysis. Each chapter consists of methodological development, extensive simulation studies, and a real data application example.

In Chapter 2, we develop a method of moments estimator for the between-study covariance matrix in random effects model multivariate meta-analysis [Chen, Manning and Dupuis, 2012]. The motivation was to solve the heterogeneity issue in the joint meta-analysis of genetic main effect and gene by environment interaction, in the context of bivariate meta-analysis. However, we note that our approach is a general statistical

method in multivariate meta-analysis, with application not limited to genetic association studies.

In Chapter 3, we combine the approaches for family data analysis and rare genetic variants analysis, and develop SKAT for quantitative traits in family samples [Chen, Meigs and Dupuis, 2013]. We expect this approach to facilitate rare genetic variants analysis in cohorts with family samples, such as the Framingham Heart Study (FHS), or individuals with cryptic relatedness.

In Chapter 4, we extend the methodology of SKAT to the context of gene by environment interaction analysis. We derive the interaction-only SKAT-type tests for both unrelated and related individuals. We hope this work will encourage investigators to study gene by environment interaction using rare genetic variants.

In Chapter 5, we summarize the findings and outline future work directions.

Chapter 2 A Method of Moments Estimator in Random Effects Multivariate Meta-Analysis

2.1 Introduction

Meta-analysis has been widely used to increase precision and power by combining studies [Cohn and Becker, 2003]. Assuming that the effect to be estimated is the same in all studies, the fixed effect meta-analysis is often successfully used to combine studies and obtain a point estimate for the effect size and its standard error. However, the underlying assumption of equal effect sizes of the fixed effect model may be violated [DerSimonian and Laird, 1986]. Different studies may come from different populations, use different protocols and have different levels of confounding or effect modifying variables. Thus, the studies may not share a common effect size. In the presence of heterogeneity, the fixed effect model underestimates the standard error of the point estimate by ignoring the between-study variance. False positive findings may be generated when the fixed effect model is inappropriately used.

The random effect model allows both within-study and between-study variances, and is more conservative than the fixed effect model in declaring significance of the effect size of interest. In the presence of heterogeneity, the random effect model provides more appropriate standard error of the point estimate and better confidence interval than the fixed effect model. When meta-analyzing a single parameter of interest, one can estimate the between-study variance by using the EM algorithm [Dempster, Laird and Rubin, 1977]

or other iterative methods; a noniterative method of moments estimator has also been proposed by DerSimonian and Laird [1986]. It was derived by equating the homogeneity test statistic to its expectation, in the presence of heterogeneity.

Meta-analysis has also been applied to two or more correlated effect estimates [Raudenbush, Becker and Kalaian, 1988], such as regression coefficients [Becker and Wu, 2007]. Analogous to the univariate case in which the inverse variance is used as the weight, the inverse of the covariance matrix is the weight in the multivariate fixed effect model. Random effect models have also been proposed to incorporate the between-study covariance matrix [Berkey et al., 1998; van Houwelingen, Arends and Stijnen, 2002; Riley et al., 2007]. However, the iterative procedure is often computer intensive and may not reach convergence. Recently, an extended DerSimonian and Laird's method of moments estimator was proposed to solve the between-study covariance matrix [Jackson, White and Thompson, 2010]. Though this method is a noniterative approach, it requires calculating each element of the matrix separately, and it is not invariant to reparametrization of effect sizes.

In this chapter, we propose a novel method of moments estimator for the between-study covariance matrix in the multivariate meta-analysis. It is also a multivariate extension of DerSimonian and Laird's univariate method of moments estimator. To our knowledge, this is the first noniterative estimator for the between-study covariance matrix in the matrix form. It is invariant to linear transformations. We perform a simulation study to

compare our method with the restricted maximum likelihood (REML) method [Jennrich and Schluchter, 1986] and Jackson's multivariate DerSimonian and Laird's method. We also apply the three random effect methods and the fixed effect approach to a real data example.

2.2 Fixed Effect Multivariate Meta-Analysis

Suppose we are interested in meta-analyzing p correlated effects from k studies. To estimate the true effect sizes

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

we need effect estimates and their covariance matrix from individual studies. For study i ($1 \leq i \leq k$), we denote the effect estimates

$$\mathbf{b}_i = \begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{ip} \end{bmatrix}$$

and their covariance matrix $\boldsymbol{\Sigma}_i$.

To meta-analyze vectors $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k$ and get a summary estimate, we need generalized least squares (GLS) methods instead of ordinary least squares (OLS) methods, because the variances of effect estimates from different studies are unequal. We first stack the k vectors to get a long vector with length kp

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_k \end{bmatrix}.$$

Assuming that the k studies are uncorrelated, we make a blockwise diagonal matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{\Sigma}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{\Sigma}_k \end{bmatrix}_{kp \times kp}.$$

This is the covariance matrix of vector \mathbf{b} .

We assume the following model holds:

$$\mathbf{b}_{kp \times 1} = \mathbf{W}_{kp \times p} \boldsymbol{\beta}_{p \times 1} + \mathbf{e}_{kp \times 1},$$

where \mathbf{W} is a stack of k identity matrices of size $p \times p$, and we assume that the error \mathbf{e} follows a multivariate normal distribution with means 0 and covariance matrix $\mathbf{\Sigma}$, which is the covariance matrix of vector \mathbf{b} .

The fixed effect model summary estimator [Raudenbush, Becker and Kalaian, 1988; Becker and Wu, 2007] is

$$\widehat{\boldsymbol{\beta}}_F = (\mathbf{W}' \mathbf{\Sigma}^{-1} \mathbf{W})^{-1} \mathbf{W}' \mathbf{\Sigma}^{-1} \mathbf{b},$$

with covariance estimator

$$Cov(\widehat{\boldsymbol{\beta}}_F) = (\mathbf{W}' \mathbf{\Sigma}^{-1} \mathbf{W})^{-1}.$$

The null hypothesis of homogeneity

$$E(\mathbf{b}_1) = E(\mathbf{b}_2) = \cdots = E(\mathbf{b}_k) = \boldsymbol{\beta}$$

can be tested using the homogeneity test statistic

$$Q = (\mathbf{b} - \mathbf{W}\widehat{\boldsymbol{\beta}}_F)' \boldsymbol{\Sigma}^{-1} (\mathbf{b} - \mathbf{W}\widehat{\boldsymbol{\beta}}_F).$$

Q is a scalar. Under the null hypothesis of no heterogeneity, it asymptotically follows a chi-square distribution with $(k - 1)p$ degrees of freedom [Becker and Wu, 2007].

2.3 Random Effect Multivariate Meta-Analysis

Similarly to the fixed effect model, we assume

$$\mathbf{b}_{kp \times 1} = \mathbf{W}_{kp \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\delta}_{kp \times 1} + \mathbf{e}_{kp \times 1}.$$

We assume that the error \mathbf{e} follows a multivariate normal distribution with means 0 and covariance matrix $\boldsymbol{\Sigma}$. The random effect vector

$$\boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_k \end{bmatrix},$$

where $\boldsymbol{\delta}_i$ ($1 \leq i \leq k$) follows a multivariate normal distribution with mean 0 and covariance matrix \mathbf{T} . Thus the covariance matrix of vector \mathbf{b} is

$$\boldsymbol{\Omega}_{kp \times kp} = \boldsymbol{\Sigma}_{kp \times kp} + \mathbf{I}_{k \times k} \otimes \mathbf{T}_{p \times p} = \begin{bmatrix} \boldsymbol{\Sigma}_1 + \mathbf{T} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_2 + \mathbf{T} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma}_k + \mathbf{T} \end{bmatrix}_{kp \times kp}.$$

$\boldsymbol{\Sigma}_i$ ($1 \leq i \leq k$) is the within-study i covariance matrix defined above, $\mathbf{I}_{k \times k}$ is the $k \times k$ identity matrix, \mathbf{T} is the between-study covariance matrix. The symbol \otimes denotes the Kronecker product of two matrices.

The random effect model summary estimator is

$$\widehat{\boldsymbol{\beta}}_R = (\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{b},$$

with covariance estimator

$$\text{Cov}(\widehat{\boldsymbol{\beta}}_R) = (\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{W})^{-1}.$$

The crucial step for the random effect approach is to estimate the between-study covariance matrix \mathbf{T} . In this chapter we use the following three methods.

2.3.1 Restricted Maximum Likelihood Method

The REML estimation can be performed by maximizing

$$L(\mathbf{T}) = -\frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\log|\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{W}| \\ - \frac{1}{2}(\mathbf{b} - \mathbf{W}(\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{b})'\boldsymbol{\Omega}^{-1}(\mathbf{b} - \mathbf{W}(\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Omega}^{-1}\mathbf{b})$$

under the constraint that \mathbf{T} is positive semi-definite [Jackson, White and Thompson, 2010; Jennrich and Schluchter, 1986].

Suppose \mathbf{T} can be written as

$$\mathbf{T} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p}\sigma_1\sigma_p & \rho_{2p}\sigma_2\sigma_p & \cdots & \sigma_p^2 \end{bmatrix}_{p \times p}.$$

The algorithm can be implemented as maximizing over $p(p+1)/2$ parameters σ_i^2, ρ_{ij}

($1 \leq i \leq p, i < j \leq p$) under constraints $\sigma_i^2 \geq 0, -1 \leq \rho_{ij} \leq 1$.

2.3.2 Jackson's Multivariate DerSimonian and Laird's Method

Jackson's multivariate DerSimonian and Laird's method estimates each entry of \mathbf{T} separately [Jackson, White and Thompson, 2010].

$$\hat{\mathbf{T}}_{i,j} = \frac{\sum_{u=1}^k \frac{(\mathbf{b}_{u(i)} - \tilde{b}_{i[j]})(\mathbf{b}_{u(j)} - \tilde{b}_{j[i]})}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}} - \sum_{u=1}^k \frac{\boldsymbol{\Sigma}_{u(i,j)}}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}} + \frac{\sum_{u=1}^k \frac{\boldsymbol{\Sigma}_{u(i,j)}}{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}}{\sum_{u=1}^k \frac{1}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}}}$$

$$\frac{\sum_{u=1}^k \frac{1}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}} - \frac{\sum_{u=1}^k \frac{1}{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}}{\sum_{u=1}^k \frac{1}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}}}$$

is the element on row i ($1 \leq i \leq p$) and column j ($1 \leq j \leq p$) of $\hat{\mathbf{T}}$, where

$$\tilde{b}_{i[j]} = \frac{\sum_{u=1}^k \frac{\mathbf{b}_{u(i)}}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}}}{\sum_{u=1}^k \frac{1}{\sqrt{\boldsymbol{\Sigma}_{u(i,i)}\boldsymbol{\Sigma}_{u(j,j)}}}},$$

and $\mathbf{b}_{u(i)}$ is the i th ($1 \leq i \leq p$) element in the vector \mathbf{b}_u ($1 \leq u \leq k$), $\boldsymbol{\Sigma}_{u(i,j)}$ is the element on row i ($1 \leq i \leq p$) and column j ($1 \leq j \leq p$) of $\boldsymbol{\Sigma}_u$ ($1 \leq u \leq k$). While the calculation is generally faster than REML, Jackson's method requires calculating p^2 weighted means $\tilde{b}_{i[j]}$ as intermediates for a $p \times p$ matrix $\hat{\mathbf{T}}$, and it is not invariant to reparametrization of effect sizes.

2.3.3 Chen's Multivariate DerSimonian and Laird's Method

Let

$$\Psi = \text{Cov}(\widehat{\boldsymbol{\beta}}_F) = (\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1} = \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1},$$

$$\Phi = \Psi^{-1} - \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} \Psi \boldsymbol{\Sigma}_i^{-1} = \sum_{i=1}^k \left[\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_i^{-1} \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1} \boldsymbol{\Sigma}_i^{-1} \right],$$

$$\mathbf{A} = \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F) (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)' - (k-1) \mathbf{I}_{p \times p},$$

where $\boldsymbol{\Sigma}_j, \mathbf{b}_j$ ($1 \leq j \leq k$) are defined above, $\widehat{\boldsymbol{\beta}}_F$ is the fixed effect estimate defined above.

Then

$$E(\mathbf{A}) = \Phi \mathbf{T}.$$

A symmetric method of moments estimator for \mathbf{T} is

$$\widehat{\mathbf{T}} = \frac{\Phi^{-1} \mathbf{A} + \mathbf{A}' \Phi^{-1}}{2}.$$

See Appendix A for the derivation.

2.3.3.1 Invariance Property to Linear Transformations

Our estimator is invariant to linear transformations. Suppose the effect sizes are transformed

$$\boldsymbol{\beta}_{new} = \mathbf{C} \boldsymbol{\beta},$$

where \mathbf{C} is an invertible $p \times p$ matrix, then

$$\mathbf{b}_{new,i} = \mathbf{C} \mathbf{b}_i,$$

$$\boldsymbol{\Sigma}_{new,i} = \mathbf{C} \boldsymbol{\Sigma}_i \mathbf{C}',$$

for $1 \leq i \leq k$, where k is the number of studies. We have

$$\boldsymbol{\Sigma}_{new,i}^{-1} = (\mathbf{C}')^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{C}^{-1},$$

$$\boldsymbol{\Psi}_{new} = \left(\sum_{j=1}^k \boldsymbol{\Sigma}_{new,j}^{-1} \right)^{-1} = \left(\sum_{j=1}^k (\mathbf{C}')^{-1} \boldsymbol{\Sigma}_j^{-1} \mathbf{C}^{-1} \right)^{-1} = \mathbf{C} \boldsymbol{\Psi} \mathbf{C}',$$

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{new,F} &= \boldsymbol{\Psi}_{new} \sum_{j=1}^k \boldsymbol{\Sigma}_{new,j}^{-1} \mathbf{b}_{new,j} = \mathbf{C} \boldsymbol{\Psi} \mathbf{C}' \sum_{j=1}^k (\mathbf{C}')^{-1} \boldsymbol{\Sigma}_j^{-1} \mathbf{C}^{-1} \mathbf{C} \mathbf{b}_j = \mathbf{C} \boldsymbol{\Psi} \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \mathbf{b}_j \\ &= \mathbf{C} \widehat{\boldsymbol{\beta}}_F, \end{aligned}$$

$$\boldsymbol{\Phi}_{new} = \boldsymbol{\Psi}_{new}^{-1} - \sum_{i=1}^k \boldsymbol{\Sigma}_{new,i}^{-1} \boldsymbol{\Psi}_{new} \boldsymbol{\Sigma}_{new,i}^{-1}$$

$$\begin{aligned} &= (\mathbf{C}')^{-1} \boldsymbol{\Psi}^{-1} \mathbf{C}^{-1} - \sum_{i=1}^k (\mathbf{C}')^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{C}^{-1} \mathbf{C} \boldsymbol{\Psi} \mathbf{C}' (\mathbf{C}')^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{C}^{-1} \\ &= (\mathbf{C}')^{-1} \boldsymbol{\Phi} \mathbf{C}^{-1}, \end{aligned}$$

$$\mathbf{A}_{new} = \sum_{j=1}^k \boldsymbol{\Sigma}_{new,j}^{-1} (\mathbf{b}_{new,j} - \widehat{\boldsymbol{\beta}}_{new,F}) (\mathbf{b}_{new,j} - \widehat{\boldsymbol{\beta}}_{new,F})' - (k-1) \mathbf{I}_{p \times p}$$

$$= \sum_{j=1}^k (\mathbf{C}')^{-1} \boldsymbol{\Sigma}_j^{-1} \mathbf{C}^{-1} \mathbf{C} (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F) (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)' \mathbf{C}' - (k-1) \mathbf{I}_{p \times p}$$

$$= (\mathbf{C}')^{-1} \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F) (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)' \mathbf{C}' - (k-1) \mathbf{I}_{p \times p}$$

$$= (\mathbf{C}')^{-1} \mathbf{A} \mathbf{C}',$$

$$\widehat{\boldsymbol{\Gamma}}_{new} = \frac{\boldsymbol{\Phi}_{new}^{-1} \mathbf{A}_{new} + \mathbf{A}_{new}' \boldsymbol{\Phi}_{new}^{-1}}{2}$$

$$= \frac{\mathbf{C} \boldsymbol{\Phi}^{-1} \mathbf{C}' (\mathbf{C}')^{-1} \mathbf{A} \mathbf{C}' + \mathbf{C} \mathbf{A}' \mathbf{C}^{-1} \mathbf{C} \boldsymbol{\Phi}^{-1} \mathbf{C}'}{2}$$

$$\begin{aligned}
&= \mathbf{C}\hat{\mathbf{T}}\mathbf{C}' , \\
\text{Cov}(\hat{\boldsymbol{\beta}}_{new,R}) &= \left(\sum_{j=1}^k (\boldsymbol{\Sigma}_{new,j} + \hat{\mathbf{T}}_{new})^{-1} \right)^{-1} = \left(\sum_{j=1}^k (\mathbf{C}(\boldsymbol{\Sigma}_j + \hat{\mathbf{T}})\mathbf{C}')^{-1} \right)^{-1} \\
&= \mathbf{C}\text{Cov}(\hat{\boldsymbol{\beta}}_R)\mathbf{C}' , \\
\hat{\boldsymbol{\beta}}_{new,R} &= \text{Cov}(\hat{\boldsymbol{\beta}}_{new,R}) \sum_{j=1}^k (\boldsymbol{\Sigma}_{new,j} + \hat{\mathbf{T}}_{new})^{-1} \mathbf{b}_{new,j} \\
&= \mathbf{C}\text{Cov}(\hat{\boldsymbol{\beta}}_R)\mathbf{C}' \sum_{j=1}^k (\mathbf{C}')^{-1} (\boldsymbol{\Sigma}_j + \hat{\mathbf{T}})^{-1} \mathbf{C}^{-1} \mathbf{C}\mathbf{b}_i \\
&= \mathbf{C}\text{Cov}(\hat{\boldsymbol{\beta}}_R) \sum_{j=1}^k (\boldsymbol{\Sigma}_j + \hat{\mathbf{T}})^{-1} \mathbf{b}_i \\
&= \mathbf{C}\hat{\boldsymbol{\beta}}_R .
\end{aligned}$$

Thus, the meta-analysis of linearly transformed effects gives the same results.

2.3.3.2 Connection with Univariate DerSimonian and Laird's Estimator

Our estimator is a multivariate extension of DerSimonian and Laird's method of moments estimator for the between-study variance. When the number of effects $p = 1$, $\boldsymbol{\Sigma}_i = s_i^2$ (variance of the effect size estimate in study i), $\mathbf{T} = \tau^2$ (between-study variance) are all scalar. Thus the homogeneity test statistic can be written as

$$Q = \sum_{j=1}^k \frac{1}{s_j^2} (b_j - \hat{\beta}_F)^2 .$$

It follows that

$$A = Q - (k - 1),$$

$$\Phi = \sum_{i=1}^k \left[s_i^{-2} - s_i^{-4} \left(\sum_{j=1}^k s_j^{-2} \right)^{-1} \right],$$

$$E(A) = \Phi \tau^2.$$

The method of moments estimator

$$\hat{\tau}^2 = \frac{Q - (k - 1)}{\Phi} = \frac{Q - (k - 1)}{\left(\sum_{i=1}^k \frac{1}{s_i^2} - \frac{\sum_{i=1}^k \frac{1}{s_i^4}}{\sum_{j=1}^k \frac{1}{s_j^2}} \right)}$$

is the same as DerSimonian and Laird's method of moments estimator.

2.3.3.3 Homogeneity Test in the Presence of Heterogeneity

In the presence of heterogeneity, the homogeneity test statistic Q no longer follows a chi-square distribution. When the between-study covariance matrix $\mathbf{T} \neq 0$, the homogeneity test statistic

$$\begin{aligned} Q &= (\mathbf{b} - \mathbf{W}\hat{\boldsymbol{\beta}}_F)' \boldsymbol{\Sigma}^{-1} (\mathbf{b} - \mathbf{W}\hat{\boldsymbol{\beta}}_F) \\ &= \sum_{j=1}^k (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F). \end{aligned}$$

Its expectation can be written as

$$\begin{aligned} E(Q) &= E \left[\sum_{j=1}^k (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F) \right] \\ &= \sum_{j=1}^k E \left\{ \text{tr} \left[(\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F) \right] \right\} \end{aligned}$$

$$\begin{aligned}
&= E \left\{ \sum_{j=1}^k \text{tr} \left[\Sigma_j^{-1} (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F) (\mathbf{b}_j - \hat{\boldsymbol{\beta}}_F)' \right] \right\} \\
&= \text{tr} [(k-1) \mathbf{I}_{p \times p} + \boldsymbol{\Phi} \mathbf{T}] \\
&= (k-1)p + \text{tr}(\boldsymbol{\Phi} \mathbf{T}).
\end{aligned}$$

Under the hypothesis of no heterogeneity, $\mathbf{T} = \mathbf{0}$, $E(Q) = (k-1)p$. This result is consistent with the expectation of a chi-square distribution with $(k-1)p$ degrees of freedom, which corresponds to the asymptotic distribution of Q in the absence of heterogeneity. It is obvious that in all circumstances the homogeneity test statistic

$$Q = (k-1)p + \text{tr}(\mathbf{A})$$

2.3.4 Positive Semi-Definiteness of Covariance Matrix Estimators

Since \mathbf{T} is a covariance matrix, it should be positive semi-definite. While we can do the maximization under the constraint that \mathbf{T} is positive semi-definite using the REML method, we generally have no guarantee that Jackson's or Chen's method of moments estimator would be positive semi-definite, especially when heterogeneity is low. This is analogous to the one dimensional case, whereas $\hat{\tau}^2$ may be negative if $Q < (k-1)$, when heterogeneity is low. One may leave a negative $\hat{\tau}^2$ as is, or set it to 0 because the variance cannot be negative. Similarly, we have two strategies for a between-study covariance matrix estimate which is not positive semi-definite: leave it as is, or make it a positive semi-definite matrix. A remedy for a covariance matrix estimate which is not positive semi-definite is discussed below, which we adopted in all the simulation studies

and real data analysis. Jackson, White and Thompson [2010] used the same strategy in their paper.

Since $\hat{\mathbf{T}}$ is symmetric, it is diagonalizable. There exists an orthogonal matrix \mathbf{U} such that

$$\mathbf{U}'\hat{\mathbf{T}}\mathbf{U} = \mathbf{D},$$

where \mathbf{D} is a diagonal matrix with all eigenvalues of $\hat{\mathbf{T}}$ on the diagonal, then

$$\hat{\mathbf{T}} = \mathbf{U}\mathbf{D}\mathbf{U}'.$$

Let \mathbf{D}_{psd} be a diagonal matrix with the same elements as \mathbf{D} , except that all negative elements on the diagonal in \mathbf{D} are set to be 0 in \mathbf{D}_{psd} . Then $\hat{\mathbf{T}}_{psd} = \mathbf{U}\mathbf{D}_{psd}\mathbf{U}'$ is positive semi-definite. However, $\hat{\mathbf{T}}_{psd}$ is a biased estimator.

2.4 Simulation Studies

2.4.1 Simulation Design

To compare the performance of the REML method, multivariate DerSimonian and Laird's method by Jackson et al. (MDLJ) and multivariate DerSimonian and Laird's method by Chen et al. (MDLC), we conducted simulation studies in the context of bivariate meta-analysis.

We considered 10 studies with different sample sizes. One hundred between-study variances were generated from a chi-square distribution with 1 df, and values less than 0.016 or greater than 2.7 were discarded (corresponding approximately to the 10% and 90% quantiles of 1 df chi-square distribution). Then we randomly chose 2 sets of 10 variances

out of the remaining values, sorted and paired them. The smallest pair was assigned to the first study as the within-study variances of the two effects and so on until the largest pair was assigned to the last study. The within-study correlation was set to 0.2 or 0.8 for all 10 studies.

We followed the procedure by Higgins and Thompson to calculate the between-study variances [Jackson, White and Thompson, 2010; Higgins and Thompson, 2002]. Since the variances for both outcomes were simulated from the same population, we first calculated the typical within-study variance

$$\sigma_{\infty}^2 = \lim_{n \rightarrow \infty} \frac{(n-1) \sum_{i=1}^n 1/\sigma_i^2}{\left(\sum_{i=1}^n 1/\sigma_i^2\right)^2 - \sum_{i=1}^n 1/\sigma_i^4},$$

where σ_i^2 was generated as discussed above. For this parameter setting we have $\sigma_{\infty}^2 = 0.15$. Then we computed the between-study variances from

$$I_1^2 = \frac{\mathbf{T}_{11}}{\mathbf{T}_{11} + \sigma_{\infty}^2} = \frac{\mathbf{T}_{11}}{\mathbf{T}_{11} + 0.15}$$

$$I_2^2 = \frac{\mathbf{T}_{22}}{\mathbf{T}_{22} + \sigma_{\infty}^2} = \frac{\mathbf{T}_{22}}{\mathbf{T}_{22} + 0.15}$$

where I_1^2 and I_2^2 are the proportions of marginal variation in the first and second effects due to heterogeneity, respectively. \mathbf{T}_{11} and \mathbf{T}_{22} are the between-study variances for the first and second effects. We considered 9 scenarios, in which each of I_1^2 and I_2^2 was set to 0.2, 0.5 or 0.8 to simulate low, moderate and high heterogeneity for each effect. The between-study correlation was set to 0.2 or 0.8 to calculate the covariance.

For each study i ($1 \leq i \leq 10$), we generated the effect size vector \mathbf{b}_i from a bivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}_i + \mathbf{T}$, where $\mathbf{\Sigma}_i$ is the within-study covariance matrix and \mathbf{T} is the between-study covariance matrix.

2.4.2 Simulation Results

We summarized the simulation results from 1000 replicates for 10 studies with between-study correlation 0.2 and within-study correlation 0.2 in Table 2.1 and Table 2.2. We only presented results for the first effect and the covariance, since the second effect shows similar results. Table 2.1 shows the bias, mean squared error (MSE) for the first summary effect estimator $\hat{\beta}_1$ and 95% confidence interval coverage for the first effect size. The confidence interval was constructed as $\hat{\beta}_1 \pm t_{0.025,9}SE(\hat{\beta}_1)$ [Follmann and Proschan, 1999]. Table 2.2 shows the bias of the between-study variance estimator for the first effect $\hat{\mathbf{T}}_{11}$ and that of the between-study covariance estimator $\hat{\mathbf{T}}_{12}$, and their corresponding mean squared errors. It also shows the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space: either at least one of the variance estimates is 0, or the absolute value of the correlation coefficient is 1. Since the between-study covariance matrix is always positive semi-definite, this column also indicates the percentage of the between-study covariance matrix not being positive definite, that is, its minimum eigenvalue is equal to 0. To allow for rounding errors, we consider values less than 10^{-8} to be equal to 0.

Table 2.1 Simulation results for summary effect estimates (between-study correlation 0.2, within-study correlation 0.2)

I_1^2	I_2^2	\mathbf{T}_{11}	\mathbf{T}_{12}	\mathbf{T}_{22}	Method	Bias $\hat{\beta}_1$	MSE $\hat{\beta}_1$	Coverage $\hat{\beta}_1$
0.2	0.2	0.0375	0.0075	0.0375	REML	-0.0065	0.0331	0.962
					MDLJ	-0.0073	0.0326	0.963
					MDLC	-0.0072	0.0326	0.963
0.2	0.5	0.0375	0.0150	0.1500	REML	-0.0001	0.0311	0.952
					MDLJ	-0.0017	0.0307	0.957
					MDLC	-0.0016	0.0307	0.958
0.2	0.8	0.0375	0.0300	0.6000	REML	0.0048	0.0334	0.962
					MDLJ	0.0054	0.0328	0.969
					MDLC	0.0058	0.0328	0.967
0.5	0.2	0.1500	0.0150	0.0375	REML	0.0103	0.0562	0.934
					MDLJ	0.0104	0.0548	0.938
					MDLC	0.0102	0.0549	0.939
0.5	0.5	0.1500	0.0300	0.1500	REML	0.0108	0.0527	0.924
					MDLJ	0.0121	0.0521	0.930
					MDLC	0.0120	0.0521	0.929
0.5	0.8	0.1500	0.0600	0.6000	REML	-0.0001	0.0499	0.925
					MDLJ	-0.0012	0.0492	0.937
					MDLC	-0.0009	0.0492	0.935
0.8	0.2	0.6000	0.0300	0.0375	REML	0.0099	0.1132	0.929
					MDLJ	0.0112	0.1130	0.933
					MDLC	0.0113	0.1131	0.933
0.8	0.5	0.6000	0.0600	0.1500	REML	0.0055	0.1147	0.925
					MDLJ	0.0046	0.1136	0.934
					MDLC	0.0045	0.1136	0.932
0.8	0.8	0.6000	0.1200	0.6000	REML	0.0256	0.1003	0.925
					MDLJ	0.0264	0.0998	0.928
					MDLC	0.0262	0.0998	0.930

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. \mathbf{T}_{11} and \mathbf{T}_{22} denote the between-study variances for the first and second effects, and \mathbf{T}_{12} denotes the between-study covariance. Coverage denotes the coverage of 95% confidence interval.

Table 2.2 Simulation results for between-study covariance matrix (between-study correlation 0.2, within-study correlation 0.2)

I_1^2	I_2^2	Method	Bias \hat{T}_{11}	Bias \hat{T}_{12}	MSE \hat{T}_{11}	MSE \hat{T}_{12}	% boundary
0.2	0.2	REML	0.0491	0.0167	0.0235	0.0078	0.888
		MDLJ	0.0473	0.0086	0.0142	0.0051	0.819
		MDLC	0.0476	0.0085	0.0142	0.0052	0.824
0.2	0.5	REML	0.0452	0.0096	0.0218	0.0118	0.838
		MDLJ	0.0447	0.0011	0.0154	0.0092	0.748
		MDLC	0.0447	0.0009	0.0152	0.0094	0.751
0.2	0.8	REML	0.0553	0.0082	0.0270	0.0343	0.708
		MDLJ	0.0499	-0.0044	0.0164	0.0262	0.629
		MDLC	0.0505	-0.0044	0.0168	0.0262	0.641
0.5	0.2	REML	0.0369	0.0185	0.0470	0.0137	0.814
		MDLJ	0.0265	0.0091	0.0361	0.0104	0.738
		MDLC	0.0267	0.0089	0.0361	0.0105	0.737
0.5	0.5	REML	0.0278	0.0051	0.0490	0.0195	0.731
		MDLJ	0.0207	-0.0017	0.0317	0.0161	0.621
		MDLC	0.0213	-0.0018	0.0315	0.0162	0.622
0.5	0.8	REML	0.0350	0.0034	0.0530	0.0500	0.552
		MDLJ	0.0265	-0.0079	0.0383	0.0454	0.454
		MDLC	0.0268	-0.0071	0.0380	0.0458	0.451
0.8	0.2	REML	0.0296	0.0125	0.2405	0.0347	0.739
		MDLJ	0.0296	-0.0009	0.2763	0.0280	0.619
		MDLC	0.0299	-0.0006	0.2764	0.0279	0.612
0.8	0.5	REML	0.0769	0.0149	0.3099	0.0567	0.545
		MDLJ	0.0504	0.0007	0.2876	0.0501	0.443
		MDLC	0.0507	0.0013	0.2879	0.0500	0.439
0.8	0.8	REML	0.0065	0.0034	0.2153	0.1108	0.288
		MDLJ	0.0033	0.0048	0.2551	0.1358	0.210
		MDLC	0.0042	0.0055	0.2559	0.1350	0.214

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. % boundary denotes the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space.

We can see from Tables 2.1 and 2.2 that all three methods give very similar results, and our new approach is closer to Jackson's method than to REML. This is not surprising because REML is a likelihood-based iterative method, while the other two are methods of moments. In all scenarios REML gives more between-study covariance matrix estimates at the boundary than the other non-iterative methods.

Bias of $\hat{\mathbf{T}}_{11}$ is generally greater than 0, because we pull back negative eigenvalues to 0 when fixing covariance matrix estimates that are not positive semi-definite, we somehow bias the diagonal elements upwards. We can see that as the heterogeneity increases, this bias generally decreases, and the proportion of between-study covariance matrix estimates at the boundary also decreases. This is consistent with our prior knowledge that fixed effect models are usually preferred when heterogeneity is low and random effect models are more appropriate when heterogeneity is high.

Simulation results from 1000 replicates for 10 studies with other correlation coefficient settings are summarized in Tables 2.3 – 2.8.

Table 2.3 Simulation results for summary effect estimates (between-study correlation 0.2, within-study correlation 0.8)

I_1^2	I_2^2	\mathbf{T}_{11}	\mathbf{T}_{12}	\mathbf{T}_{22}	Method	Bias $\hat{\beta}_1$	MSE $\hat{\beta}_1$	Coverage $\hat{\beta}_1$
0.2	0.2	0.0375	0.0075	0.0375	REML	-0.0065	0.0331	0.962
					MDLJ	-0.0073	0.0326	0.963
					MDLC	-0.0072	0.0326	0.963
0.2	0.5	0.0375	0.0150	0.1500	REML	-0.0001	0.0311	0.952
					MDLJ	-0.0017	0.0307	0.957
					MDLC	-0.0016	0.0307	0.958
0.2	0.8	0.0375	0.0300	0.6000	REML	0.0048	0.0334	0.962
					MDLJ	0.0054	0.0328	0.969
					MDLC	0.0058	0.0328	0.967
0.5	0.2	0.1500	0.0150	0.0375	REML	0.0103	0.0562	0.934
					MDLJ	0.0104	0.0548	0.938
					MDLC	0.0102	0.0549	0.939
0.5	0.5	0.1500	0.0300	0.1500	REML	0.0108	0.0527	0.924
					MDLJ	0.0121	0.0521	0.930
					MDLC	0.0120	0.0521	0.929
0.5	0.8	0.1500	0.0600	0.6000	REML	-0.0001	0.0499	0.925
					MDLJ	-0.0012	0.0492	0.937
					MDLC	-0.0009	0.0492	0.935
0.8	0.2	0.6000	0.0300	0.0375	REML	0.0099	0.1132	0.929
					MDLJ	0.0112	0.1130	0.933
					MDLC	0.0113	0.1131	0.933
0.8	0.5	0.6000	0.0600	0.1500	REML	0.0055	0.1147	0.925
					MDLJ	0.0046	0.1136	0.934
					MDLC	0.0045	0.1136	0.932
0.8	0.8	0.6000	0.1200	0.6000	REML	0.0256	0.1003	0.925
					MDLJ	0.0264	0.0998	0.928
					MDLC	0.0262	0.0998	0.930

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. \mathbf{T}_{11} and \mathbf{T}_{22} denote the between-study variances for the first and second effects, and \mathbf{T}_{12} denotes the between-study covariance. Coverage denotes the coverage of 95% confidence interval.

Table 2.4 Simulation results for between-study covariance matrix (between-study correlation 0.2, within-study correlation 0.8)

I_1^2	I_2^2	Method	Bias \hat{T}_{11}	Bias \hat{T}_{12}	MSE \hat{T}_{11}	MSE \hat{T}_{12}	% boundary
0.2	0.2	REML	0.0473	0.0396	0.0159	0.0106	0.819
		MDLJ	0.0426	0.0281	0.0134	0.0090	0.752
		MDLC	0.0418	0.0261	0.0125	0.0081	0.752
0.2	0.5	REML	0.0467	0.0363	0.0192	0.0149	0.765
		MDLJ	0.0431	0.0188	0.0136	0.0130	0.659
		MDLC	0.0438	0.0201	0.0141	0.0133	0.658
0.2	0.8	REML	0.0516	0.0584	0.0219	0.0388	0.670
		MDLJ	0.0541	0.0148	0.0149	0.0333	0.549
		MDLC	0.0635	0.0256	0.0196	0.0438	0.562
0.5	0.2	REML	0.0532	0.0502	0.0536	0.0224	0.740
		MDLJ	0.0330	0.0280	0.0412	0.0156	0.663
		MDLC	0.0308	0.0272	0.0388	0.0152	0.652
0.5	0.5	REML	0.0442	0.0387	0.0498	0.0265	0.623
		MDLJ	0.0376	0.0202	0.0414	0.0228	0.518
		MDLC	0.0386	0.0221	0.0412	0.0222	0.517
0.5	0.8	REML	0.0384	0.0403	0.0557	0.0540	0.506
		MDLJ	0.0272	0.0088	0.0346	0.0495	0.389
		MDLC	0.0347	0.0109	0.0410	0.0563	0.409
0.8	0.2	REML	0.0601	0.0493	0.2386	0.0364	0.728
		MDLJ	0.0140	0.0040	0.2838	0.0298	0.576
		MDLC	0.0198	0.0086	0.3190	0.0322	0.598
0.8	0.5	REML	0.0215	0.0388	0.2269	0.0551	0.492
		MDLJ	-0.0110	0.0094	0.2642	0.0503	0.400
		MDLC	-0.0117	0.0098	0.2764	0.0534	0.393
0.8	0.8	REML	0.0118	0.0138	0.2125	0.1221	0.280
		MDLJ	-0.0106	-0.0176	0.2189	0.1241	0.203
		MDLC	-0.0033	-0.0101	0.2391	0.1295	0.247

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. % boundary denotes the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space.

Table 2.5 Simulation results for summary effect estimates (between-study correlation 0.8, within-study correlation 0.2)

I_1^2	I_2^2	\mathbf{T}_{11}	\mathbf{T}_{12}	\mathbf{T}_{22}	Method	Bias $\hat{\beta}_1$	MSE $\hat{\beta}_1$	Coverage $\hat{\beta}_1$
0.2	0.2	0.0375	0.0300	0.0375	REML	-0.0000	0.0308	0.958
					MDLJ	0.0009	0.0310	0.959
					MDLC	0.0008	0.0310	0.958
0.2	0.5	0.0375	0.0600	0.1500	REML	-0.0066	0.0325	0.956
					MDLJ	-0.0053	0.0327	0.956
					MDLC	-0.0052	0.0326	0.957
0.2	0.8	0.0375	0.1200	0.6000	REML	0.0065	0.0319	0.960
					MDLJ	0.0067	0.0317	0.967
					MDLC	0.0065	0.0317	0.967
0.5	0.2	0.1500	0.0600	0.0375	REML	-0.0082	0.0495	0.934
					MDLJ	-0.0099	0.0493	0.951
					MDLC	-0.0099	0.0492	0.952
0.5	0.5	0.1500	0.1200	0.1500	REML	-0.0049	0.0484	0.950
					MDLJ	-0.0044	0.0484	0.951
					MDLC	-0.0046	0.0485	0.953
0.5	0.8	0.1500	0.2400	0.6000	REML	-0.0075	0.0494	0.941
					MDLJ	-0.0075	0.0496	0.943
					MDLC	-0.0074	0.0495	0.945
0.8	0.2	0.6000	0.1200	0.0375	REML	-0.0024	0.1075	0.932
					MDLJ	-0.0003	0.1081	0.930
					MDLC	-0.0001	0.1083	0.931
0.8	0.5	0.6000	0.2400	0.1500	REML	-0.0070	0.1008	0.926
					MDLJ	-0.0073	0.1006	0.925
					MDLC	-0.0074	0.1005	0.926
0.8	0.8	0.6000	0.4800	0.6000	REML	-0.0121	0.1094	0.936
					MDLJ	-0.0136	0.1087	0.943
					MDLC	-0.0134	0.1087	0.941

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. \mathbf{T}_{11} and \mathbf{T}_{22} denote the between-study variances for the first and second effects, and \mathbf{T}_{12} denotes the between-study covariance. Coverage denotes the coverage of 95% confidence interval.

Table 2.6 Simulation results for between-study covariance matrix (between-study correlation 0.8, within-study correlation 0.2)

I_1^2	I_2^2	Method	Bias \hat{T}_{11}	Bias \hat{T}_{12}	MSE \hat{T}_{11}	MSE \hat{T}_{12}	% boundary
0.2	0.2	REML	0.0523	0.0064	0.0343	0.0103	0.915
		MDLJ	0.0497	-0.0015	0.0173	0.0062	0.859
		MDLC	0.0499	-0.0016	0.0173	0.0062	0.860
0.2	0.5	REML	0.0509	0.0041	0.0208	0.0134	0.845
		MDLJ	0.0507	-0.0057	0.0157	0.0114	0.768
		MDLC	0.0512	-0.0060	0.0158	0.0116	0.781
0.2	0.8	REML	0.0532	0.0074	0.0228	0.0374	0.779
		MDLJ	0.0510	-0.0054	0.0152	0.0300	0.694
		MDLC	0.0515	-0.0056	0.0152	0.0305	0.698
0.5	0.2	REML	0.0333	0.0000	0.0511	0.0132	0.866
		MDLJ	0.0306	-0.0090	0.0406	0.0114	0.792
		MDLC	0.0307	-0.0089	0.0405	0.0116	0.790
0.5	0.5	REML	0.0364	-0.0059	0.0419	0.0227	0.771
		MDLJ	0.0301	-0.0167	0.0328	0.0192	0.682
		MDLC	0.0305	-0.0167	0.0328	0.0194	0.675
0.5	0.8	REML	0.0422	-0.0034	0.0457	0.0628	0.677
		MDLJ	0.0343	-0.0139	0.0356	0.0560	0.624
		MDLC	0.0344	-0.0134	0.0351	0.0561	0.631
0.8	0.2	REML	0.0145	-0.0011	0.2381	0.0330	0.802
		MDLJ	0.0132	-0.0077	0.2815	0.0328	0.691
		MDLC	0.0130	-0.0077	0.2813	0.0328	0.694
0.8	0.5	REML	-0.0048	-0.0263	0.2141	0.0558	0.691
		MDLJ	-0.0000	-0.0273	0.2529	0.0649	0.597
		MDLC	-0.0004	-0.0275	0.2509	0.0642	0.598
0.8	0.8	REML	0.0283	-0.0067	0.2290	0.1503	0.511
		MDLJ	0.0227	-0.0103	0.2466	0.1845	0.437
		MDLC	0.0225	-0.0111	0.2459	0.1834	0.425

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. % boundary denotes the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space.

Table 2.7 Simulation results for summary effect estimates (between-study correlation 0.8, within-study correlation 0.8)

I_1^2	I_2^2	\mathbf{T}_{11}	\mathbf{T}_{12}	\mathbf{T}_{22}	Method	Bias $\hat{\beta}_1$	MSE $\hat{\beta}_1$	Coverage $\hat{\beta}_1$
0.2	0.2	0.0375	0.0300	0.0375	REML	0.0066	0.0317	0.952
					MDLJ	0.0062	0.0307	0.949
					MDLC	0.0057	0.0306	0.952
0.2	0.5	0.0375	0.0600	0.1500	REML	0.0095	0.0322	0.952
					MDLJ	0.0077	0.0325	0.954
					MDLC	0.0085	0.0327	0.948
0.2	0.8	0.0375	0.1200	0.6000	REML	-0.0054	0.0303	0.966
					MDLJ	-0.0081	0.0309	0.955
					MDLC	-0.0081	0.0315	0.958
0.5	0.2	0.1500	0.0600	0.0375	REML	-0.0129	0.0428	0.946
					MDLJ	-0.0126	0.0417	0.946
					MDLC	-0.0119	0.0420	0.946
0.5	0.5	0.1500	0.1200	0.1500	REML	0.0037	0.0483	0.936
					MDLJ	0.0049	0.0477	0.931
					MDLC	0.0037	0.0479	0.938
0.5	0.8	0.1500	0.2400	0.6000	REML	-0.0047	0.0502	0.930
					MDLJ	-0.0051	0.0506	0.927
					MDLC	-0.0037	0.0518	0.917
0.8	0.2	0.6000	0.1200	0.0375	REML	-0.0258	0.1043	0.931
					MDLJ	-0.0257	0.1050	0.927
					MDLC	-0.0269	0.1053	0.920
0.8	0.5	0.6000	0.2400	0.1500	REML	-0.0031	0.1126	0.925
					MDLJ	-0.0041	0.1131	0.918
					MDLC	-0.0045	0.1128	0.920
0.8	0.8	0.6000	0.4800	0.6000	REML	0.0008	0.1025	0.926
					MDLJ	0.0004	0.1041	0.923
					MDLC	-0.0003	0.1038	0.922

I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. \mathbf{T}_{11} and \mathbf{T}_{22} denote the between-study variances for the first and second effects, and \mathbf{T}_{12} denotes the between-study covariance. Coverage denotes the coverage of 95% confidence interval.

Table 2.8 Simulation results for between-study covariance matrix (between-study correlation 0.8, within-study correlation 0.8)

I_1^2	I_2^2	Method	Bias $\hat{\mathbf{T}}_{11}$	Bias $\hat{\mathbf{T}}_{12}$	MSE $\hat{\mathbf{T}}_{11}$	MSE $\hat{\mathbf{T}}_{12}$	% boundary
0.2	0.2	REML	0.0488	0.0355	0.0219	0.0137	0.899
		MDLJ	0.0427	0.0248	0.0147	0.0102	0.840
		MDLC	0.0409	0.0217	0.0136	0.0093	0.851
0.2	0.5	REML	0.0406	0.0230	0.0156	0.0150	0.825
		MDLJ	0.0349	0.0057	0.0111	0.0132	0.743
		MDLC	0.0355	0.0071	0.0114	0.0137	0.750
0.2	0.8	REML	0.0533	0.0409	0.0217	0.0428	0.748
		MDLJ	0.0449	0.0006	0.0151	0.0349	0.639
		MDLC	0.0483	0.0001	0.0157	0.0399	0.626
0.5	0.2	REML	0.0330	0.0365	0.0431	0.0215	0.811
		MDLJ	0.0198	0.0160	0.0300	0.0137	0.731
		MDLC	0.0174	0.0150	0.0302	0.0139	0.734
0.5	0.5	REML	0.0296	0.0226	0.0379	0.0267	0.723
		MDLJ	0.0205	0.0062	0.0327	0.0255	0.644
		MDLC	0.0189	0.0071	0.0321	0.0246	0.627
0.5	0.8	REML	0.0445	0.0209	0.0572	0.0692	0.575
		MDLJ	0.0329	-0.0000	0.0434	0.0725	0.507
		MDLC	0.0349	0.0011	0.0510	0.0839	0.500
0.8	0.2	REML	0.0467	0.0506	0.2187	0.0499	0.732
		MDLJ	0.0334	0.0175	0.2852	0.0420	0.631
		MDLC	0.0439	0.0294	0.3144	0.0531	0.645
0.8	0.5	REML	0.0187	0.0226	0.2210	0.0731	0.592
		MDLJ	0.0164	0.0038	0.2807	0.0769	0.491
		MDLC	0.0194	0.0067	0.2913	0.0827	0.486
0.8	0.8	REML	0.0222	0.0225	0.2453	0.2030	0.306
		MDLJ	0.0194	0.0099	0.2795	0.2294	0.261
		MDLC	0.0126	0.0054	0.2773	0.2288	0.237

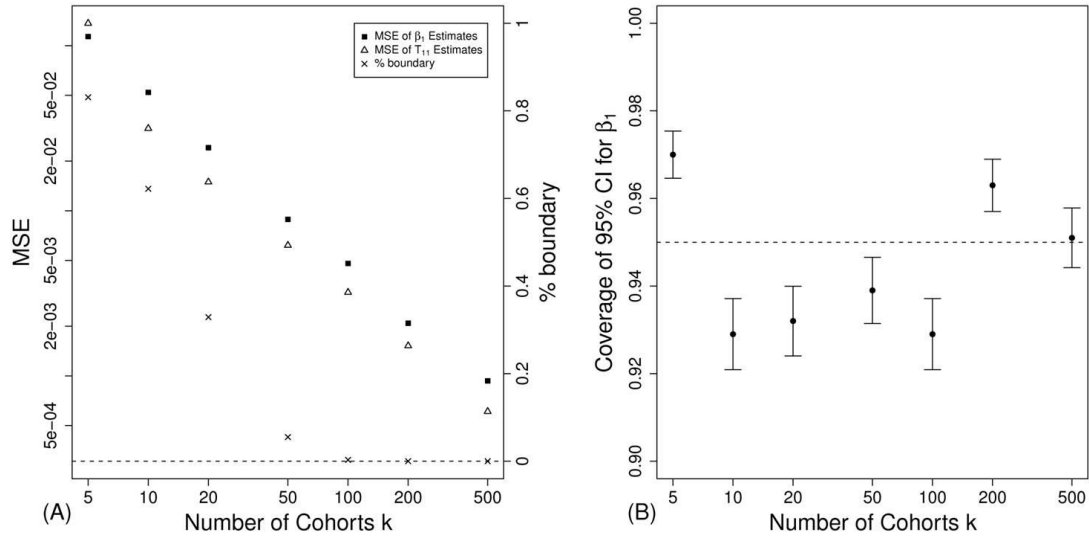
I_1^2 and I_2^2 denote the proportions of marginal variation in the first and second effects due to heterogeneity. % boundary denotes the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space.

2.4.3 Additional Simulation Studies

We performed 6 additional simulation studies to investigate the effect of the number of studies on the results. We set I_1^2 and I_2^2 to be 0.5, between-study and within-study correlation coefficients to be 0.2, and let the number of studies change from 10 to 5, 20, 50, 100, 200 and 500. In each scenario, we analyzed 1000 replicates. Since all three methods give similar results, only results using our new method are shown in Figure 2.1 and Table 2.9. As the number of studies increases, mean squared errors of $\hat{\beta}_1$, the summary effect estimate for the first effect, and $\hat{\mathbf{T}}_{11}$, the between-study variance estimate for the first effect, decrease. This is reasonable because the large-sample properties of estimators in a meta-analysis depend on the number of studies. As the number of studies goes to infinity, mean squared errors converge to 0.

We can also see that as the number of studies increases, the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space (matrix with minimum eigenvalue 0) decreases dramatically, even though the heterogeneity is only moderate. As we fix fewer and fewer covariance matrices that are not positive semi-definite, bias of $\hat{\mathbf{T}}_{11}$ also decreases quickly to near 0. There is no clear relationship between the coverage of 95% confidence interval for β_1 and the number of studies, although only the largest sample of 500 studies has the correct coverage.

Figure 2.1 Simulation results from various numbers of studies (proportions of marginal variation due to heterogeneity 0.5, between-study correlation 0.2, within-study correlation 0.2)



(A) Mean squared errors of $\hat{\beta}_1$, the first summary effect estimate, and $\hat{\tau}_{11}$, the between-study variance estimate for the first effect, and proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space. (B) Coverage of 95% confidence interval for β_1 , the first effect size.

Table 2.9 Simulation results from various numbers of studies (proportions of marginal variation due to heterogeneity 0.5, between-study correlation 0.2, within-study correlation 0.2)

k	Bias $\hat{\beta}_1$	MSE $\hat{\beta}_1$	Coverage $\hat{\beta}_1$	Bias \hat{T}_{11}	Bias \hat{T}_{12}	MSE \hat{T}_{11}	MSE \hat{T}_{12}	% boundary
5	-0.0064	0.1135	0.970	0.0989	0.0023	0.1366	0.0513	0.831
10	0.0120	0.0521	0.929	0.0213	-0.0018	0.0315	0.0162	0.622
20	0.0031	0.0241	0.932	0.0004	0.0022	0.0149	0.0082	0.329
50	0.0001	0.0089	0.939	-0.0006	0.0002	0.0062	0.0030	0.055
100	-0.0022	0.0048	0.929	0.0012	0.0003	0.0032	0.0016	0.003
200	0.0004	0.0021	0.963	0.0002	0.0003	0.0015	0.0008	0.000
500	0.0005	0.0009	0.951	-0.0001	-0.0009	0.0006	0.0003	0.000

k denotes the number of studies in the meta-analysis. Coverage denotes the coverage of 95% confidence interval. % boundary denotes the proportion of the between-study covariance matrix estimate lying on the boundary of its parameter space.

2.5 Application

In this application we use publicly available data from the base year of the High School Longitudinal Study of 2009 (HSLs:09) [Ingels et al., 2011] to illustrate our new method as a general multivariate meta-analysis approach. HSLs:09 is a nationally representative, longitudinal study of more than 21,000 9th graders in 944 schools who will be followed through their secondary and postsecondary years. We are interested in testing whether sex, socio-economic status and sex by socio-economic status interaction are predictive of the mathematics standardized theta score. We estimate the regression coefficients in each of the 8 race groups and perform multivariate meta-analyses on the regression coefficients to obtain the summary effect estimates.

Within each race group i , our model is

$$Y_{ij} = \beta_{i0} + \beta_{i1}X_{ij1} + \beta_{i2}X_{ij2} + \beta_{i3}X_{ij1}X_{ij2} + \varepsilon_{ij}$$

where Y_{ij} is the mathematics standardized theta score, X_{ij1} is the sex, coded as 1 for males and 0 for females, X_{ij2} is the socio-economic status score for student j . ε_{ij} is the normally distributed error. The regression results are summarized in Table 2.10.

We use both the fixed effects meta-analysis (FEMA) and random effects models to meta-analyze the regression coefficients from the 8 race groups. Table 2.11 shows the meta-analysis results. For this data, the homogeneity test statistic Q is 54.6, which asymptotically follows a chi-square distribution with 21 degrees of freedom under the null hypothesis of no heterogeneity. The p-value of the homogeneity test is 8.1×10^{-5} . Thus, the assumption of homogeneous effect sizes for the fixed effect model is violated. Since the fixed effect model does not take between-study variance into consideration, it greatly underestimates the covariance matrix for the summary effect estimates, resulting in an inflated Wald test statistic for testing the hypotheses $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ versus H_1 : at least one of $\beta_1, \beta_2, \beta_3$ is not 0, where $\beta_1, \beta_2, \beta_3$ are the summary effect sizes for sex, socio-economic status score and sex by socio-economic status score interaction, respectively.

Jackson's multivariate DerSimonian and Laird's method of moments and our method give close results, while the restricted maximum likelihood method yields a different

between-study covariance matrix estimate. However, all three random effect methods give very close summary effect size estimates, and the Wald test statistic reduces dramatically, compared to that from the fixed effect model.

Table 2.10 Regression results from 8 race groups

Race i	N_i	b_{i1}	b_{i2}	b_{i3}	$Cov(\mathbf{b})$		
1	163	0.3161	7.4015	0.4278	2.3568	-1.2105	0.8524
						9.7029	-6.1753
							4.4114
2	1672	-0.3201	6.9426	-0.9816	0.2529	0.1498	-0.1019
						0.7016	-0.4167
							0.2743
3	2218	0.6983	4.6680	-0.2415	0.1444	-0.0652	0.0433
						0.6481	-0.3899
							0.2608
4	204	3.2736	4.3080	0.2052	3.8428	-4.5587	3.2892
						10.3517	-6.6684
							4.8268
5	3311	-0.1599	5.6398	-0.6782	0.1161	-0.0992	0.0645
						0.4363	-0.2610
							0.1733
6	1912	-0.6989	6.3158	-0.7918	0.1603	0.0242	-0.0129
						0.7697	-0.4686
							0.3180
7	110	-3.6094	9.3429	-2.8711	3.2054	-1.1984	0.8437
						17.8889	-10.7697
							7.2101
8	11854	0.2172	6.4078	-0.6093	0.0278	0.0136	-0.0091
						0.1184	-0.0716
							0.0482

N_i is the sample size in race group i . b_{i1} is the regression coefficient for sex, b_{i2} is the regression coefficient for socio-economic status score, b_{i3} is the regression coefficient for sex by socio-economic status score interaction in race group i .

Table 2.11 Meta-analysis results for 8 race groups

Method	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)	$\hat{\mathbf{T}}$			Wald statistic
FEMA	0.0788 (0.1208)	6.2031 (0.2448)	-0.6590 (0.1550)	0	0	0	4141
REML	-0.0244 (0.2427)	6.1674 (0.4843)	-0.6679 (0.1865)	0.2090	-0.2946 1.0028	0.1172 -0.2567 0.0657	431
MDLJ	-0.0612 (0.2599)	6.1873 (0.2973)	-0.7038 (0.1888)	0.2558	-0.1221 0.1279	0.0097 0.0542 0.0501	567
MDLC	-0.0604 (0.2684)	6.1821 (0.2887)	-0.7009 (0.1894)	0.2805	-0.0948 0.1024	0.0030 0.0602 0.0532	571

$\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ denote the summary effect size estimates for sex, socio-economic status score and sex by socio-economic status score interaction, respectively. $\hat{\mathbf{T}}_{11}$, $\hat{\mathbf{T}}_{22}$, $\hat{\mathbf{T}}_{33}$, $\hat{\mathbf{T}}_{12}$, $\hat{\mathbf{T}}_{13}$, $\hat{\mathbf{T}}_{23}$ are corresponding elements of the matrix $\hat{\mathbf{T}}$, the between-study covariance matrix estimate. Wald statistic is for testing the hypotheses $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ versus H_1 : at least one of $\beta_1, \beta_2, \beta_3$ is not 0.

2.6 Discussion

We propose a new method of moments estimator for the between-study covariance matrix in the random effect multivariate meta-analysis. We have shown in our simulation studies that our method gives similar results to existing random effect model multivariate meta-analysis methods. Furthermore, our method and Jackson's multivariate DerSimonian and Laird's method give very close results in both simulation studies and real data analysis.

Our estimator is the first matrix form method of moments estimator for the between-study covariance matrix in the random effect model multivariate meta-analysis. It is invariant to linear transformations. As a non-iterative estimator, it is very easy to calculate.

Despite its long history in combining published analysis results, meta-analysis is also very useful in multi-center or multi-ethnic studies, when different cohorts can share results from the same analysis but it is often not feasible to share original data. Generally, the fixed effect model is the first choice in a meta-analysis as it is easier to calculate and interpret, and it is more powerful than random effect models. However, in the presence of heterogeneity, results from the fixed effect model are not valid, and random effect models are preferred. When performing the random effect meta-analysis for p effects, Jackson's method requires calculating p^2 weighted means $\tilde{b}_{i[j]}$ ($1 \leq i \leq p, 1 \leq j \leq p$) in intermediate steps, which are not related to corresponding fixed effect summary estimates. Specifically, the weighted means $\tilde{b}_{i[i]}$ ($1 \leq i \leq p$) are the fixed effect summary estimates in univariate meta-analyses, instead of the multivariate meta-analysis we desire. In contrast, our method directly uses the fixed effect summary estimates vector to calculate the between-study covariance matrix estimate. It does not require performing p^2 additional univariate meta-analyses to calculate the intermediates. We hope our computationally easy estimator with nice mathematical properties will help boost multivariate meta-analysis using the random effects model.

Chapter 3 Sequence Kernel Association Test for Quantitative Traits in Family Samples

3.1 Introduction

In recent years, with the advances in whole-genome sequencing technology, assessing the association of rare genetic variants with complex diseases and quantitative traits has become of great interest. Rare genetic variants may account for some of the unexplained heritability unexplained by genetic loci identified by genome-wide association studies (GWAS) [Eichler et al., 2010], because single variant tests used in GWAS are underpowered for rare genetic variants [Li and Leal, 2008]. To increase power, burden tests have been proposed [Li and Leal, 2008; Morgenthaler and Thilly, 2007; Madsen and Browning, 2009; Morris and Zeggini, 2010]; these tests are based on collapsing rare genetic variants in a predefined genomic region with either a rare variant indicator or a weighted score. These methods are most powerful when all rare genetic variants in the region have the same direction of effect and even the same effect size. Alternatively, the data-adaptive sum test [Han and Pan, 2010] and step-up approach [Hoffmann, Marini and Witte, 2010] do not make any implicit assumptions about the directions of effects and use the signs from single marker tests to determine them, but both of these approaches require permutation to evaluate statistical significance.

In contrast with burden tests, the sequence kernel association test (SKAT) [Wu et al., 2011] is a flexible and computationally efficient regression-based approach for rare

genetic variants analysis. No assumptions about the directions of effects or the effect sizes of rare genetic variants in the region are required for SKAT. Instead of requiring permutation for the p-value computation, Davies' method [Davies, 1980] is used to compute the p-values analytically for SKAT. SKAT has been shown to be much more powerful than traditional burden tests in many different scenarios. SKAT can be used in the association analysis of both dichotomous and continuous phenotypes.

Family-based study designs have been widely used in linkage analysis of diseases and quantitative traits [Falk and Rubinstein, 1987; Ott, 1989; Terwilliger and Ott, 1992; Spielman, McGinnis and Ewens, 1993]. In GWAS, ordinary regression approaches are not applicable to family data, because inflated type I error is observed when familial correlation is not appropriately modeled. For quantitative traits, instead of ordinary linear regressions, linear mixed effects models that take familial correlation as a random effect with covariance proportional to the kinship matrix is commonly used for single marker tests in GWAS [Amos, 1994; Almasy and Blangero, 1998; Pankratz, de Andrade and Therneau, 2005]. However, burden tests and other methods for joint analysis of rare genetic variants in family samples have not been well established.

In this chapter, we use the framework of linear mixed effects models to extend SKAT for rare genetic variants association analysis with quantitative traits in family data. The family-sample SKAT (famSKAT) has a different form of test statistic and distribution under the null hypothesis, but has the same underlying principle as SKAT. When there is

no familial correlation, famSKAT is equivalent to SKAT. P-values for famSKAT are also calculated analytically without requiring permutation.

We demonstrate in our simulation studies that SKAT has inflated type I error in family samples when familial correlation is not appropriately considered. By contrast, famSKAT does not suffer from this issue and has correct type I error. We also show that famSKAT is more powerful than applying SKAT to an unrelated subset of the sample. For mixed datasets with both unrelated and related individuals, as the proportion of unrelated individuals decreases, the difference in power between SKAT and famSKAT increases, with famSKAT being always the more powerful approach of the two. Thus, by using famSKAT there is no need to reduce sample size by selecting an unrelated subset of individuals. Finally, we illustrate our approach by assessing the association between rare genetic variants using glycemic traits in the Framingham Heart Study.

3.2 Burden Test for Quantitative Traits in Family Samples

Assuming a sample size of n , let the $n \times 1$ vector of the quantitative trait \mathbf{y} follow a linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where \mathbf{X} is an $n \times p$ covariate matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector consisting of fixed effects parameters (an intercept and $p - 1$ coefficients for covariates), \mathbf{G} is an $n \times q$ genotype matrix for q rare genetic variants of interest, \mathbf{W} is the pre-specified diagonal weight matrix for the rare variants of $q \times q$, $\boldsymbol{\gamma}$ is a $q \times 1$ vector $\boldsymbol{\gamma}\mathbf{1}$, $\boldsymbol{\delta}$ is an $n \times 1$ vector for the

random effects of familial correlation, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the error. The vector of error $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\delta}$ are assumed normally distributed and uncorrelated with each other:

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi}),$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n),$$

where $\boldsymbol{\Phi}$ is twice the kinship matrix of size $n \times n$ obtained from family information only, σ_G^2, σ_E^2 are corresponding variance component parameters. In this parameter setting, we are interested in testing $H_0: \boldsymbol{\gamma} = \mathbf{0}$ versus $H_1: \boldsymbol{\gamma} \neq \mathbf{0}$. This is a burden test, because we implicitly assume that all rare variants in this test share the same effect size (after weighting).

3.3 Sequence Kernel Association Test for Quantitative Traits in Family Samples

We follow the same notations as in Section 3.2, but now $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of rare variants. The vector of error $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are assumed normally distributed and uncorrelated with each other:

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \tau \mathbf{I}_q),$$

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi}),$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n),$$

where $\tau, \sigma_G^2, \sigma_E^2$ are corresponding variance component parameters. In this parameter setting, we are interested in testing $H_0: \tau = 0$ versus $H_1: \tau > 0$, which is equivalent to testing $H_0: \boldsymbol{\gamma} = \mathbf{0}$ versus $H_1: \boldsymbol{\gamma} \neq \mathbf{0}$. This is a variance component score test in the linear mixed effects model, which is a locally most powerful test [Wu et al., 2011; Lin, 1997].

Under these assumptions, the phenotypic variance can be written as

$$\text{Var}(\mathbf{y}) = \tau \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' + \sigma_G^2 \mathbf{\Phi} + \sigma_E^2 \mathbf{I}_n = \mathbf{\Sigma}.$$

The log likelihood for the linear mixed effects model is

$$l = C - \frac{1}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

To derive a score test for $H_0: \tau = 0$, we first take the derivative with respect to τ to get

$$\frac{dl}{d\tau} = -\frac{1}{2} \text{tr}(\mathbf{\Sigma}^{-1} \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}') + \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Sigma}^{-1} \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

If we use the restricted maximum likelihood instead of the maximum likelihood method, we would get a different first term, but the same second term. In both cases, if we replace $\mathbf{\Sigma}$ by its consistent estimator, and treat genotype matrix \mathbf{G} as fixed, then the first term in the score function is fixed and independent of phenotype data \mathbf{y} . Following the same rationale used in the derivation of the SKAT score statistic [Liu, Lin and Ghosh, 2007; Kwee et al., 2008], we take twice the second term to be derived as our test statistic.

Under the null hypothesis $\tau = 0$, we can estimate

$$\hat{\mathbf{\Sigma}} = \hat{\sigma}_G^2 \mathbf{\Phi} + \hat{\sigma}_E^2 \mathbf{I}_n,$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \hat{\mathbf{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{\Sigma}}^{-1} \mathbf{y}$$

by fitting the null linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

The maximum likelihood estimators can be obtained using the function `lmekin` from R package `kinship`. We replace $\boldsymbol{\beta}$, σ_G^2 and σ_E^2 (and hence $\boldsymbol{\Sigma}$) by their maximum likelihood estimators and take

$$Q = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{G} \mathbf{W} \mathbf{W}' \mathbf{G}' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

as the famSKAT test statistic. Under the null hypothesis, the variance of the residuals is

$$\text{Var}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Sigma}} - \mathbf{X}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}_0.$$

Thus

$$Q \sim \sum_{i=1}^q \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix $\mathbf{W}\mathbf{G}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_0\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}$. The p-value can be computed analytically by Davies' method [Davies, 1980] or Kuonen's saddlepoint method [Kuonen, 1999].

3.3.1 Connection with SKAT in Unrelated Individuals

We note that even though the null model, test statistic, residual variance and null distribution of famSKAT have different forms compared to those of SKAT, they are directly connected. Actually, if we add a restriction $\sigma_G^2 = 0$ to the model, famSKAT is equivalent to SKAT. Then

$$\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_E^2 \mathbf{I}_n,$$

where $\hat{\sigma}_E^2$ is estimated from the null linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

and famSKAT statistic becomes

$$Q = \frac{1}{\hat{\sigma}_E^4} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

with distribution under the null hypothesis

$$Q \sim \sum_{i=1}^q \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix $\hat{\sigma}_E^{-4} \mathbf{W}\mathbf{G}'\mathbf{P}_0\mathbf{G}\mathbf{W}$. The famSKAT statistic and its null distribution matrix are proportional to SKAT statistic and null distribution matrix with the coefficient $\hat{\sigma}_E^{-4}$. We favor this form of null distribution matrix, rather than the form proposed in Wu et al. [2011], because usually the sample size n is larger than the number of genetic variants of interest q , the non-zero eigenvalues of $\mathbf{W}\mathbf{G}'\mathbf{P}_0\mathbf{G}\mathbf{W}$ and $\mathbf{P}_0^{\frac{1}{2}}\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}'\mathbf{P}_0^{\frac{1}{2}}$ are the same, but the first matrix is of size $q \times q$, while the second matrix is of size $n \times n$.

3.3.2 Reparametrization

We note that famSKAT can also be used when we want to provide a known heritability coefficient h^2 externally, rather than estimating it from the data. By the reparametrization

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2},$$

$$\sigma^2 = \sigma_G^2 + \sigma_E^2,$$

when h^2 is known, we can use the generalized least square method to estimate only σ^2 under the null model. Then we can follow the rest of the famSKAT procedure to perform the test.

3.4 Simulation Studies

3.4.1 Type I Error

3.4.1.1 Simulation Design

To evaluate the type I error, we performed several simulation studies under the null hypothesis of no genetic association. We compared four approaches: famSKAT, family-sample burden test (famBT), unrelated-sample SKAT (unrSKAT) which only takes the unrelated subset of the sample, and SKAT. We used Kuonen's saddlepoint method [Kuonen, 1999] to compute the p-values for famSKAT, unrSKAT and SKAT.

We set the heritability of the trait

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} = 0.5 .$$

For each parameter setting, we simulated 100 genotype datasets with a total sample size of 1000 and 20 single nucleotide polymorphisms (SNP) with minor allele frequency (MAF) in the founders randomly sampled from a uniform distribution of 0.005 to 0.05, and with low ($r = 0.1$), moderate ($r = 0.5$), or high ($r = 0.7$) linkage disequilibrium (LD) between adjacent SNPs in the founders. The correlation between farther SNPs decays as an autoregressive model with order 1. We simulated haplotypes for unrelated founders with desired MAF and LD structure using the same procedure as HapSim [Montana, 2005], then we passed down the haplotypes to the next generation to simulate sib pairs, and took the remaining founders as unrelated individuals. Thus we created

genotype datasets mixed with unrelated individuals and sib pairs, and let the proportion of unrelated individuals decrease from 75% to 50%, 25%, 0%. For each genotype dataset, 10,000 phenotype datasets including covariates were simulated by using the model

$$\mathbf{y} = 0.05\mathbf{age} + 0.5\mathbf{sex} + \boldsymbol{\varepsilon},$$

where \mathbf{age} is a vector of continuous values generated from a normal distribution with mean 50 and standard deviation 5, \mathbf{sex} is a vector of dichotomous values generated from a Bernoulli distribution with probability 0.5, $\boldsymbol{\varepsilon}$ follows a multivariate normal distribution with means 0 and covariance matrix $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = h^2\boldsymbol{\Phi} + (1 - h^2)\mathbf{I}_n.$$

We calculated the p-values of famSKAT, famBT, unrSKAT and SKAT using the Wu weights [Wu et al., 2011], corresponding to the square of a beta density function of the observed MAF in the founders with parameters 1 and 25. We computed the empirical type I error at α levels of 0.01, 0.001 and 0.0001 by counting the proportion of p-values less than or equal to the corresponding α level in the 1 million genotype-phenotype datasets.

3.4.1.2 Simulation Results

Table 3.1 shows the empirical type I errors of famSKAT, famBT, unrSKAT and SKAT at different α levels in 3 LD scenarios and 4 scenarios for the proportion of unrelated individuals. The results suggest that when SKAT is directly applied to the full sample with correlated individuals, it has inflated type I error at all α levels. The empirical type I error tends to be higher when LD decays. In contrast, famSKAT, famBT and unrSKAT

retain the correct type I errors. Thus, in subsequent power simulations we only investigated these three approaches. The distributions of the p-values from the four approaches for the scenario of LD between adjacent SNPs $r = 0.5$ and proportion of unrelated individuals 0% are shown in Figure 3.1. We found that famSKAT, famBT and unrSKAT all had uniform distribution of the p-values, while the distribution of the p-values from SKAT was more likely to be small, explaining the inflated type I error.

Figure 3.1 Distribution of null p-values of famSKAT, famBT, unrSKAT and SKAT

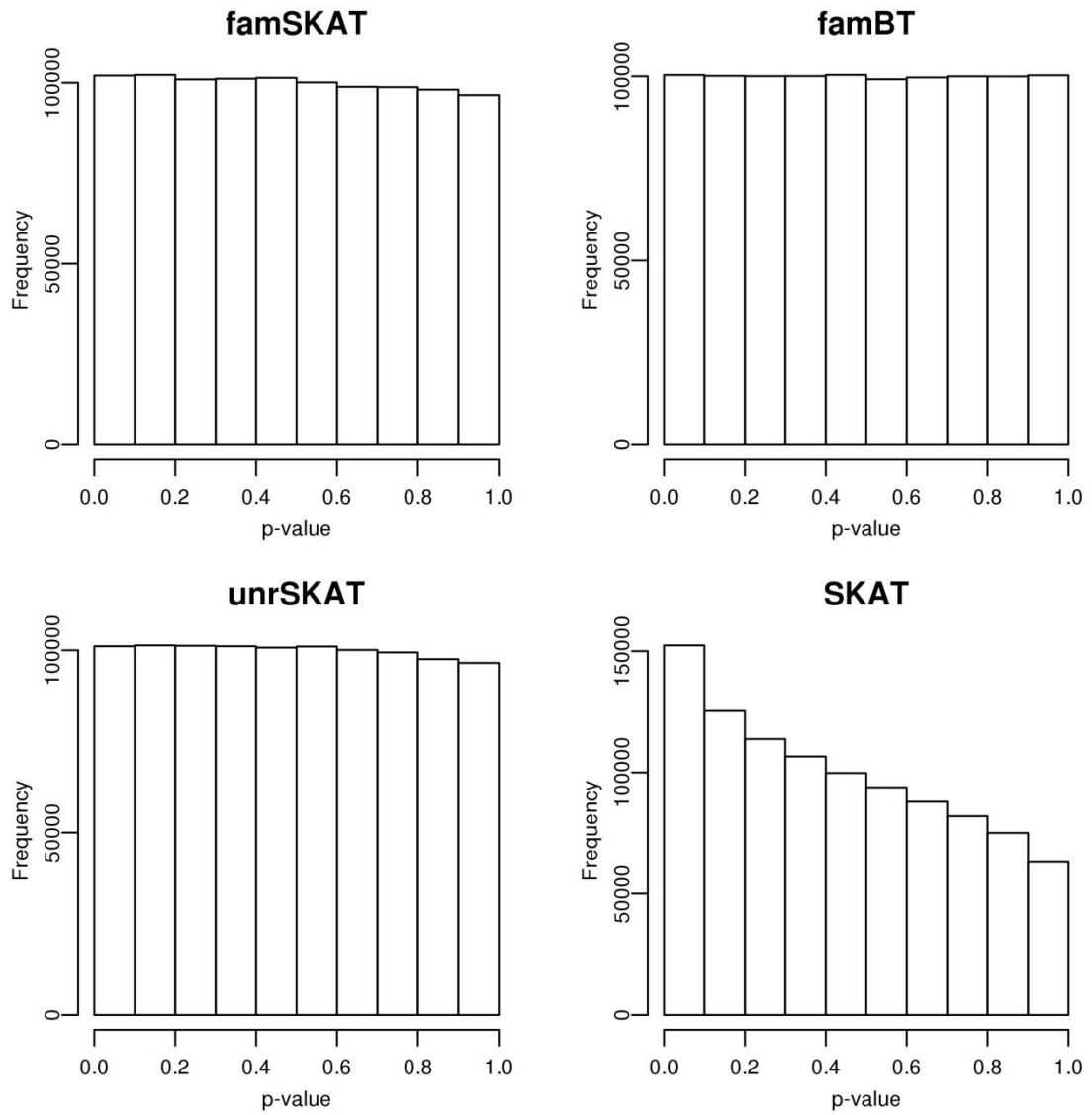


Table 3.1 Type I errors of famSKAT, famBT, unrSKAT and SKAT

LD	Unrelated%	α level	famSKAT	famBT	unrSKAT	SKAT
r = 0.1	0%	0.01	0.00970	0.01005	0.00871	0.02734
		0.001	0.00096	0.00100	0.00080	0.00383
		0.0001	0.00009	0.00009	0.00008	0.00052
	25%	0.01	0.00963	0.01026	0.00893	0.02168
		0.001	0.00090	0.00106	0.00078	0.00279
		0.0001	0.00009	0.00012	0.00008	0.00036
	50%	0.01	0.00954	0.01007	0.00901	0.01685
		0.001	0.00094	0.00101	0.00087	0.00200
		0.0001	0.00009	0.00009	0.00008	0.00024
	75%	0.01	0.00965	0.01034	0.00935	0.01276
		0.001	0.00089	0.00107	0.00088	0.00133
		0.0001	0.00009	0.00012	0.00009	0.00016
r = 0.5	0%	0.01	0.00998	0.00999	0.00966	0.01991
		0.001	0.00097	0.00101	0.00088	0.00244
		0.0001	0.00008	0.00008	0.00009	0.00029
	25%	0.01	0.01003	0.01011	0.00975	0.01699
		0.001	0.00097	0.00099	0.00092	0.00202
		0.0001	0.00010	0.00011	0.00008	0.00025
	50%	0.01	0.00999	0.01005	0.00982	0.01439
		0.001	0.00096	0.00103	0.00095	0.00161
		0.0001	0.00009	0.00011	0.00009	0.00018
	75%	0.01	0.01012	0.01008	0.00991	0.01200
		0.001	0.00098	0.00101	0.00094	0.00124
		0.0001	0.00009	0.00011	0.00009	0.00013
r = 0.7	0%	0.01	0.00984	0.01003	0.00968	0.01687
		0.001	0.00092	0.00106	0.00092	0.00197
		0.0001	0.00010	0.00011	0.00009	0.00025
	25%	0.01	0.00972	0.00997	0.00962	0.01471
		0.001	0.00091	0.00099	0.00087	0.00171
		0.0001	0.00010	0.00011	0.00007	0.00019
	50%	0.01	0.00978	0.01002	0.00952	0.01290
		0.001	0.00092	0.00101	0.00092	0.00139
		0.0001	0.00009	0.00009	0.00010	0.00015
	75%	0.01	0.00997	0.01009	0.00989	0.01161
		0.001	0.00096	0.00101	0.00094	0.00120
		0.0001	0.00011	0.00011	0.00009	0.00013

Empirical type I errors were calculated as the proportion of p-values less than or equal to the corresponding α level in 1 million genotype-phenotype datasets.

3.4.2 Power

3.4.2.1 Simulation Design

To evaluate the power of famSKAT, famBT and unrSKAT, we set the heritability of phenotype $h^2 = 0.5$ and LD between adjacent SNPs in the founders $r = 0.5$, and performed simulations under different scenarios. For each parameter setting, we simulated 100 genotype datasets with a total sample size 1000 and 20 SNPs with MAF in the founders randomly sampled from a uniform distribution of 0.005 to 0.05. Similar to the null simulation setting, we simulated genotype datasets mixed with unrelated individuals and sib pairs, and changed the proportion of unrelated individuals from 75% to 50%, 25%, 0%. For each genotype dataset \mathbf{G} , 10,000 phenotype datasets including covariates were simulated by using the model

$$\mathbf{y} = 0.05\mathbf{age} + 0.5\mathbf{sex} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where \mathbf{age} , \mathbf{sex} and $\boldsymbol{\varepsilon}$ are generated in the same way as in the type I error simulations, $\boldsymbol{\gamma}$ is a vector consisting of the effect sizes of the causal SNPs. We varied the proportion of causal SNPs from 20% to 50% and 80%, and we simulated both same and opposite directions of effects. Causal SNPs were randomly selected out of the 20 SNPs for each phenotype replicate, and in each parameter setting the effect sizes of causal SNPs were determined by

$$\gamma_i = \sqrt{\frac{c}{2MAF_i(1 - MAF_i)}},$$

where MAF_i is the MAF used to generate the genotype dataset for causal SNP i , and c is a constant for all causal SNPs in each phenotype replicate, calculated as

$$c = \frac{R^2}{\mathbf{v}'\mathbf{D}\mathbf{v}},$$

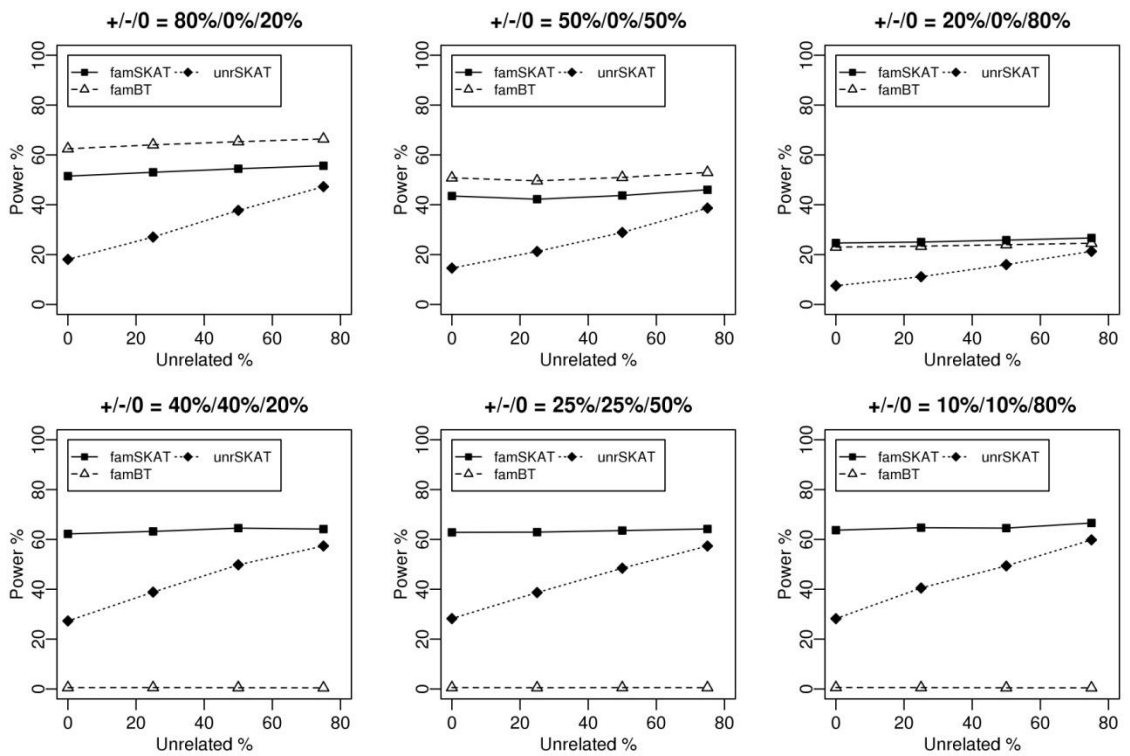
where R^2 , the total proportion of variance explained by all causal SNPs, was fixed at 1% for scenarios when all causal SNPs had effects in the same direction, and 5% for scenarios when 50% of the causal SNPs had positive effects and 50% had negative effects. \mathbf{D} is the LD correlation matrix for the 20 SNPs, and \mathbf{v} is a vector indicating the directions of causal SNP effects in each replicate. We used the same weights for famSKAT, unrSKAT and famBT, which were the Wu weights calculated from the observed MAF in founders. The empirical power was evaluated at the α level of 0.001.

3.4.2.2 Simulation Results

Power simulation results of famSKAT, famBT and unrSKAT are shown in Figure 3.2. In all scenarios, 20 SNPs were analyzed. We simulated scenarios in which the proportion of causal SNPs was 20%, 50% or 80%, with effects in the same or opposite directions. As the proportion of unrelated individuals decreases from 75% to 50%, 25% and 0%, the sample size for unrSKAT also decreases from 875 to 750, 625 and 500, respectively. As a result, the power of unrSKAT also drops. In contrast, the power of famSKAT and famBT remains almost constant, regardless of the proportion of unrelated individuals. FamBT has higher power than famSKAT when the proportion of causal SNPs is greater than or equal to 50% and all causal SNPs have the same direction of effects, but it has almost no power in scenarios when causal SNPs have opposite directions of effects. Generally, famSKAT performs well in all these scenarios, suggesting that famSKAT is an omnibus

method which does not have compromised power for different mixtures of related and unrelated individuals.

Figure 3.2 Power comparisons of famSKAT, famBT and unrSKAT



Empirical power calculated at α level of 0.001. The sample consists of sib pairs and unrelated individuals. The total sample size in each scenario is 1000, and the total number of SNPs analyzed is 20. In each panel, +/-/0 indicates the proportion of SNPs with positive effects, negative effects and no effects.

3.5 Analysis of Framingham Heart Study Data

3.5.1 Candidate Gene Study

We used genotype data from Framingham SNP Health Association Resource (SHARe) and phenotype data from the Framingham Heart Study to analyze the association with two glycemic traits: fasting glucose and log-transformed fasting insulin. We restricted our analyses to SNPs with MAF less than 5% within 100kb of 16 gene regions selected for their prior association with fasting glucose, and 2 genes reported to be associated with log-transformed fasting insulin [Dupuis et al., 2010]. We adjusted the fasting glucose analysis for age and sex, and log-transformed insulin was additionally adjusted for body mass index (BMI). We performed famSKAT and famBT for all individuals with both genotype and phenotype available, and performed SKAT for only a subset of unrelated individuals. For comparison purpose, we calculated the MAF using a subset of unrelated individuals and used Wu weights for all three methods.

We investigated the association between fasting glucose and rare genetic variants in 16 gene regions previously shown to be associated in large scale GWAS [Dupuis et al., 2010]: *ADCY5*, *ADRA2A*, *C2CD4B*, *CRY2*, *DGKB-AGMO*, *FADS1*, *G6PC2*, *GCK*, *GCKR*, *GLIS3*, *MADD*, *MTNR1B*, *PROX1*, *SLC2A2*, *SLC30A8*, and *TCF7L2*. The results are shown in Table 3.2 and Table 3.3. After adjusting for multiple testing using a Bonferroni correction, we detected no association between fasting glucose and rare genetic variants in the selected gene regions at the family-wise α level of 0.05, for all three methods. *CRY2* reaches the nominal significance level with a p-value of 0.0381

using famSKAT and 0.0085 using famBT, and *G6PC2* reaches the nominal significance level with a p-value of 0.0418 using famSKAT, but none of these gene regions reaches nominal statistical significance when evaluated using unrSKAT.

We also investigated the association between log-transformed fasting insulin and rare genetic variants in 2 gene regions previously shown to be associated in large scale GWAS [Dupuis et al., 2010]: *GCKR* and *IGF1*. After adjusting for multiple testing using a Bonferroni correction, *IGF1* shows association with log transformed fasting insulin with a nominal p-value of 0.0232 using famSKAT and 0.0234 using famBT. Neither gene reaches even the nominal significance level using SKAT.

Tables 3.2 and 3.3 also show that the sample size for unrSKAT is much smaller than that for famSKAT and famBT, because there are many families in the study even though the Framingham Heart Study is not a family-based cohort. Thus, by selecting unrelated individuals we greatly reduced the sample size. Because some SNPs with rare minor alleles may not be polymorphic in the subset of unrelated individuals, for some gene regions the number of SNPs for unrSKAT is smaller than the number of SNPs for famSKAT and famBT.

Table 3.2 Candidate gene study results from unrSKAT

Gene	Chr	Start	Stop	unrSKAT		
				Sample Size	N SNPs	p-value
Trait: fasting glucose						
<i>ADCY5</i>	3	124486089	124650082	1924	17	0.9698
<i>ADRA2A</i>	10	112826911	112830560	1924	5	0.7293
<i>C2CD4B</i>	15	60243029	60244774	1924	3	0.7104
<i>CRY2</i>	11	45825533	45861375	1924	7	0.7299
<i>DGKB-AGMO</i>	7	14153770	15568165	1924	72	0.1992
<i>FADS1</i>	11	61323677	61340886	1924	5	0.7049
<i>G6PC2</i>	2	169465996	169474756	1924	22	0.1373
<i>GCK</i>	7	44150395	44165412	1924	4	0.2486
<i>GCKR</i>	2	27573210	27600054	1924	6	0.0930
<i>GLIS3</i>	9	3814128	4290035	1924	56	0.9822
<i>MADD</i>	11	47247775	47308158	1924	7	0.6316
<i>MTNR1B</i>	11	92342437	92355596	1924	11	0.0833
<i>PROX1</i>	1	212228483	212276385	1924	33	0.1082
<i>SLC2A2</i>	3	172196831	172227462	1924	5	0.6836
<i>SLC30A8</i>	8	118216518	118258134	1924	12	0.7869
<i>TCF7L2</i>	10	114699999	114916060	1924	7	0.2147
Trait: log-transformed fasting insulin						
<i>GCKR</i>	2	27573210	27600054	1840	6	0.1016
<i>IGF1</i>	12	101335584	101398508	1840	16	0.1258

Association analysis with fasting glucose and log-transformed fasting insulin from the

Framingham Heart Study and genotype data from the Framingham SNP Health

Association Resource, using unrSKAT. Fasting glucose was adjusted for age and sex, and

log-transformed fasting insulin was adjusted for age, sex and body mass index. SNPs

with MAF less than 5% within 100kb of each gene region were included in the analysis.

Gene and SNP locations were reported on National Center for Biotechnology Information

(NCBI) build 36. Only a subset of unrelated participants with available genotypes and

phenotype were analyzed using unrSKAT. Wu weights with beta (1, 25) based on the

MAF in a subset of unrelated individuals were used.

Table 3.3 Candidate gene study results from famSKAT and famBT

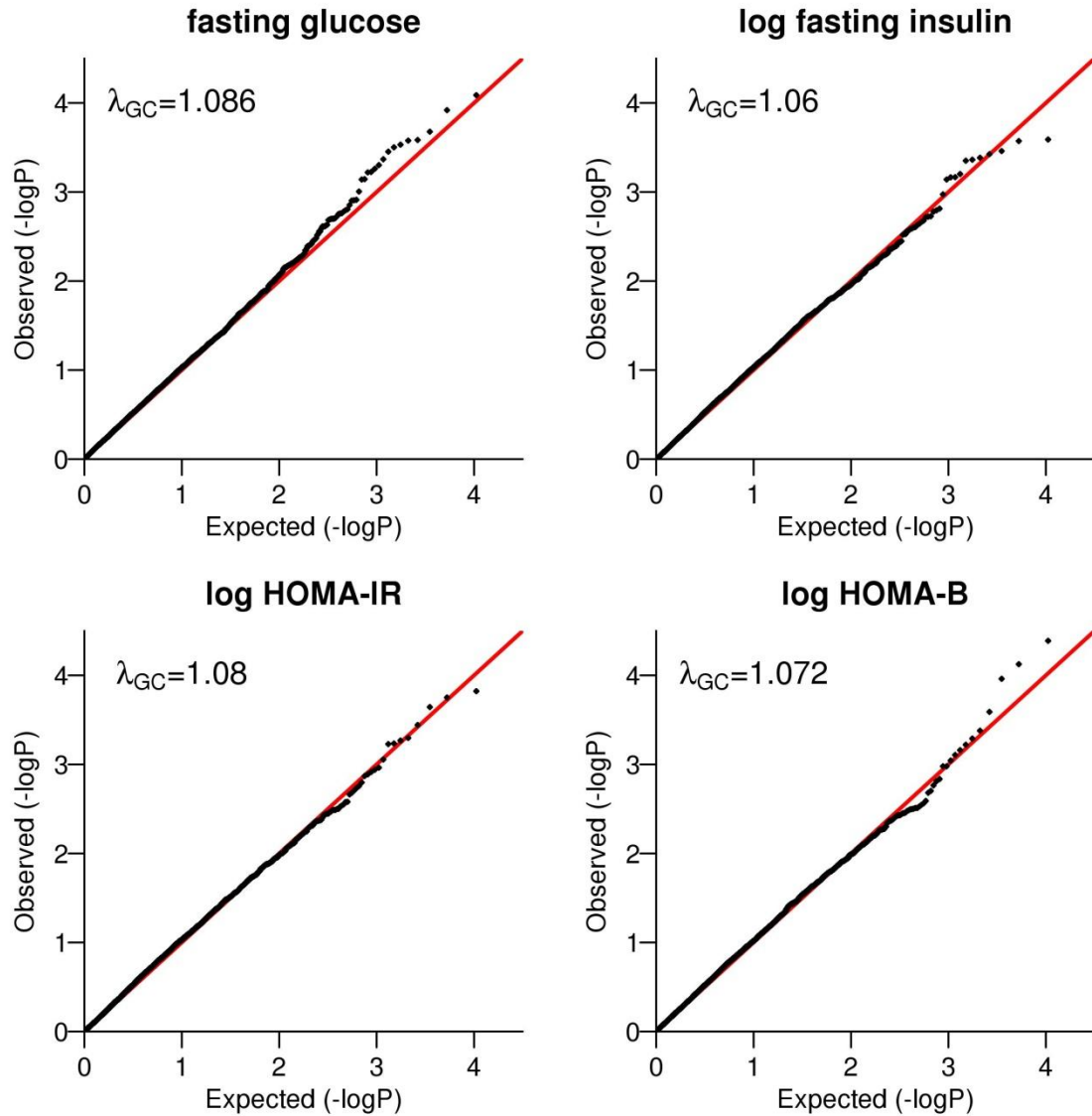
Gene	famSKAT			famBT		
	Sample Size	N SNPs	p-value	Sample Size	N SNPs	p-value
Trait: fasting glucose						
<i>ADCY5</i>	6479	18	0.9125	6479	18	0.5842
<i>ADRA2A</i>	6479	5	0.9557	6479	5	0.9453
<i>C2CD4B</i>	6479	3	0.9576	6479	3	0.8209
<i>CRY2</i>	6479	8	0.0381	6479	8	0.0085
<i>DGKB-AGMO</i>	6479	76	0.6579	6479	76	0.3648
<i>FADS1</i>	6479	5	0.5245	6479	5	0.3723
<i>G6PC2</i>	6479	24	0.0418	6479	24	0.1173
<i>GCK</i>	6479	4	0.6283	6479	4	0.5864
<i>GCKR</i>	6479	7	0.2603	6479	7	0.2461
<i>GLIS3</i>	6479	57	0.9170	6479	57	0.3920
<i>MADD</i>	6479	7	0.6571	6479	7	0.5145
<i>MTNR1B</i>	6479	11	0.8384	6479	11	0.7243
<i>PROX1</i>	6479	34	0.2414	6479	34	0.3976
<i>SLC2A2</i>	6479	5	0.8383	6479	5	0.7365
<i>SLC30A8</i>	6479	12	0.1766	6479	12	0.0970
<i>TCF7L2</i>	6479	7	0.1035	6479	7	0.4249
Trait: log-transformed fasting insulin						
<i>GCKR</i>	6031	7	0.4878	6031	7	0.1586
<i>IGF1</i>	6031	16	0.0232	6031	16	0.0234

Association analysis with fasting glucose and log-transformed fasting insulin from the Framingham Heart Study and genotype data from the Framingham SNP Health Association Resource, using famSKAT and famBT. Fasting glucose was adjusted for age and sex, and log-transformed fasting insulin was adjusted for age, sex and body mass index. SNPs with MAF less than 5% within 100kb of each gene region were included in the analysis. All individuals with available genotypes and phenotype were analyzed with famSKAT and famBT. Wu weights with beta (1, 25) based on the MAF in a subset of unrelated individuals were used.

3.5.2 Sliding Window Analysis

We also performed a genome-wide sliding window analysis on these two traits, as well as log-transformed homeostatic model assessment for insulin resistance (HOMA-IR) and homeostatic model assessment for β -cell function (HOMA-B), using SHARe genotype data. We only included SNPs with MAF less than 5% and ran the analysis using a sliding window of 500kb, with 250kb overlap each with previous and subsequent windows. We removed windows with 0 or 1 SNP, resulting in 10,546 windows for all autosomes with the number of SNPs ranging from 2 to 76 with a median of 18. No window reached the genome-wide significance using famSKAT, famBT or unrSKAT. The Quantile-Quantile plots for famSKAT are shown in Figure 3.3. There is minimal inflation of the p-values from this genome-wide analysis.

Figure 3.3 Quantile-Quantile plots for famSKAT in the genome-wide sliding window analysis on four glyceemic traits

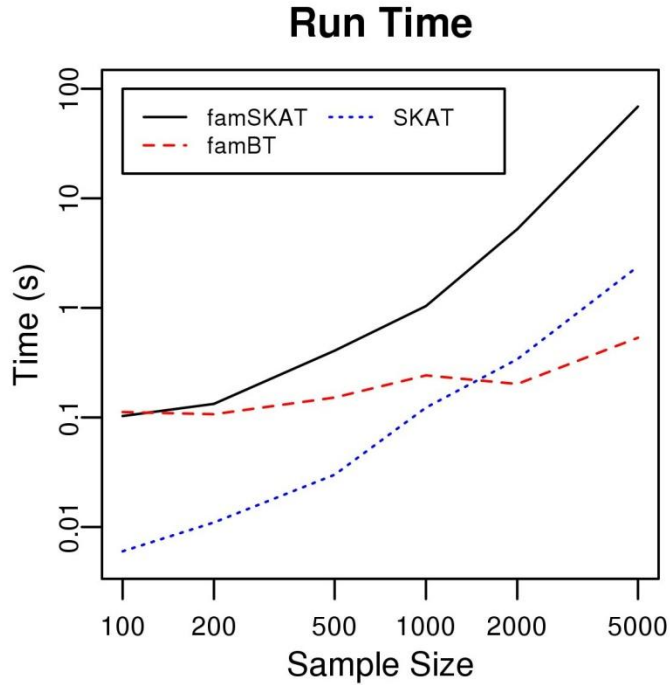


The p-values were plotted as minus log base 10 p-values. The genomic control (GC) factor λ_{GC} was computed as the ratio of median chi-square statistics with 1 df corresponding to observed and expected p-values.

3.6 Computation Time

The computation time of famSKAT depends on both the sample size and the number of SNPs. The empirical run time of famSKAT, famBT and SKAT in analyzing sib pairs with indicated total sample sizes on a single computing node with 2.33 GHz central processing unit (CPU) and 4 GB memory is shown in Figure 3.4. With a small sample size, the limiting step in famSKAT is fitting the null linear mixed effects model, so the computation time is comparable to that of famBT, which also requires fitting a linear mixed effects model. As the sample size increases, all three methods require more computation time, and the time of famSKAT and SKAT increases dramatically. Both famSKAT and SKAT require matrix calculation, and the limiting step in famSKAT becomes inverting the matrix $\hat{\Sigma}$, which takes about 90% of the computation time when the sample size is 5000. The genome-wide sliding window analysis of SHARe genotype data using a sliding window of 500kb takes about 5 hours for chromosome 1 on a single computing node with 2.33 GHz CPU and 4 GB memory.

Figure 3.4 Run time of famSKAT, famBT and SKAT in analyzing 20 SNPs



3.7 Discussion

In this chapter, we propose famSKAT as an extension of SKAT which can be applied to data with familial correlation. We demonstrate that famSKAT is a general and flexible variance component score test approach, which is equivalent to SKAT when the familial variance component is set to 0. It can be applied to quantitative traits with unknown or known heritability.

Compared with famBT, famSKAT is advantageous in power when the proportion of causal SNPs in a genomic region is small, and when not all causal SNPs have the same

direction of effects. As expected, famBT outperforms famSKAT when the proportion of causal SNPs is greater than or equal to 50% and all these SNPs have positive effects, but the performance of famSKAT in these scenarios is still satisfactory. In real data analysis, when we do not have sufficient a priori information about the proportion of causal SNPs or the directions of effects, famSKAT would be a better choice over famBT.

We show that when SKAT is inappropriately applied to correlated data, it has inflated type I error. Thus, the best we can do for SKAT is to select unrelated individuals from the whole sample. However, our power simulations demonstrate that this strategy reduces power in many scenarios. In contrast, we do not need to reduce our sample size if we use famSKAT. Our real data example from the Framingham Heart Study also shows that SKAT does not even have an observation which reaches the nominal significance level of 0.05.

Common genetic variants at 16 gene regions chosen for fasting glucose and 2 gene regions chosen for log-transformed fasting insulin have been shown to be associated with either trait in large GWAS [Dupuis et al., 2010]. However, we do not have solid evidence to show that there is strong association between either trait and the rare genetic variants in these regions. We noticed that the sample size in this analysis was far smaller than in Dupuis et al. [2010], which reduced power. In addition, the SHARe project was not specifically designed for rare variants analysis, so most SNPs in our genotype dataset are common SNPs and were excluded from the analysis. With the progress of sequencing

studies, we should be able to identify many more rare variants and perform a candidate gene or even genome-wide analysis again using the new genotype dataset with dense rare genetic variants. On the other hand, some gene regions may be truly associated with the trait only through common SNPs, so we do not expect to identify the association with rare genetic variants for all these gene regions we selected.

With the development in sequencing technology and decreasing cost, sequencing data which contain a lot of rare genetic variants have become available, not only for case-control studies, but also for cohorts that include family members. Based on SKAT, one of the most powerful rare genetic variants analysis methods to date, we developed famSKAT in the hope of facilitating rare genetic variants analysis to identify novel genes associated with quantitative traits. With famSKAT, cohorts with family data can perform the association analysis with rare genetic variants, using as much data as possible, without having to select unrelated individuals from the pedigree.

For calculating the p-values, we recommend using Kuonen's saddlepoint method [Kuonen, 1999] instead of Davies' method [Davies, 1980]. As a method based on numerical integration, Davies' method requires specifying the accuracy. When the p-value is expected to be very small, Davies' method cannot calculate it accurately. Table 3.4 shows this numerical issue in a power simulation context. Davies' method suffers from negative and zero p-values (and possibly significant round-off error) regardless of the accuracy specified. In contrast, Kuonen's method does not have such issues. Thus, if

we perform a genome-wide rare variants analysis using sequence data, from which we expect extreme low p-values, Kuonen's method is a better choice than Davies' method.

Even though famSKAT was developed for analyzing rare genetic variants, it can also be used for common variant analysis, combined common and rare variant analysis or conditional association analyses. Depending on the research hypothesis, common variants can be treated as fixed effects in the model, or random effects along with the rare genetic variants. Recently, Schifano et al. [2012] developed a SNP set association analysis approach for common variants analysis in family data, which is essentially equivalent to our method. The use of famSKAT combined with the collapsing of some very rare genetic variants such as singletons is also possible. Similar with SKAT, external weights based on annotation information or functional prediction can be incorporated to further boost power.

Table 3.4 Comparison of Kuonen's and Davies' methods in calculating p-values in the tail

Method	Accuracy	Minimum p-value	Median p-value	Maximum p-value	% p-value < 0	% p-value = 0	% round-off error [§]
Kuonen	NA	1.2×10^{-23}	2.1×10^{-9}	0.021	0%	0%	0%
Davies	10^{-4}	-1.2×10^{-6}	0	0.022	0.87%	83.31%	0%
	10^{-5}	-5.6×10^{-7}	0	0.022	6.04%	71.37%	0%
	10^{-6}	-3.1×10^{-8}	0	0.022	12.14%	54.32%	0%
	10^{-7}	-4.4×10^{-9}	0	0.022	16.38%	38.19%	0%
	10^{-8}	-2.4×10^{-9}	0	0.022	27.65%	23.15%	0%
	10^{-9}	-2.0×10^{-10}	1.8×10^{-9}	0.022	21.41%	14.86%	0%
	10^{-10}	-3.3×10^{-11}	1.9×10^{-9}	0.022	17.89%	9.23%	0%
	10^{-11}	-5.7×10^{-13}	1.9×10^{-9}	0.022	7.64%	5.38%	0%
	10^{-12}	-9.7×10^{-14}	1.9×10^{-9}	0.022	5.93%	2.93%	0%
	10^{-13}	-2.4×10^{-14}	1.9×10^{-9}	0.022	4.79%	1.53%	0%
	10^{-14}	-4.4×10^{-16}	1.9×10^{-9}	0.022	0.01%	0.89%	20.34%
	10^{-15}	-2.9×10^{-15}	1.9×10^{-9}	0.022	1.40%	0.58%	99.67%

[§] Proportion of significant round-off error in the calculation, returned by the function

davies from R package CompQuadForm. Using our power simulation framework, we simulated a scenario with phenotype heritability h^2 equal to 0.5, LD between adjacent SNPs in the founders r set to 0.5. We simulated 500 sib pairs and 20 SNPs with founders' MAF randomly sampled from a uniform distribution of 0.005 to 0.05. Of these 20 SNPs, 16 were neutral and 4 were positively associated with the trait, explaining 5% of the phenotypic variance in total. We analyzed 10000 replicates using famSKAT with Kuonen's method, and Davies' method with accuracy from 10^{-4} to 10^{-15} .

Chapter 4 Methods for Rare Genetic Variants Analysis in Detecting Gene by Environment Interaction on Quantitative Traits

4.1 Introduction

Traditional genome-wide association studies (GWAS) have been successfully applied to identify a large number of genetic markers associated with complex diseases and related quantitative traits. However, for most complex diseases and quantitative traits, all genetic markers identified so far only explain a small proportion of heritability in these traits, suggesting that a lot of genetic determinants are still undiscovered. The most commonly used approach in traditional GWAS is the single marker test of common genetic variants. Eichler et al. [2010] suggested that gene by environment interaction and rare genetic variants may both account for some of the unexplained heritability.

Statistical methods for detecting gene by environment interaction have been well established in the context of common genetic variants [Kraft et al., 2007; Manning et al., 2011]. To determine if a common genetic variant interacts with an environmental variable, which is also included in the regression model as a covariate, we can either assess the interaction term alone, or jointly test both the genetic main effect and gene by environment interaction terms. By using the first approach, we are usually interested in testing the gene by environment interaction, regardless of the presence of a significant genetic main effect. However, by using the second approach, we are testing if the genetic marker is associated with the trait of interest, allowing for gene by environmental

covariate interaction. These methods combined with multivariate meta-analysis have led to the discovery of novel common loci associated with fasting insulin [Manning et al., 2012].

On the other hand, rare genetic variants analysis has become a popular research field in genetic association studies, and many statistical methods for rare variants analysis have been proposed [Han and Pan, 2010; Hoffmann, Marini and Witte, 2010; Li and Leal, 2008; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Wu et al., 2011]. Of these methods, SKAT [Wu et al., 2011] has been shown as one of the most powerful approach in various scenarios. However, all of these methods focus on the main effect association analysis of rare genetic variants.

Compared with common variants analysis, rare variants analysis often requires a larger sample size to attain comparable power. Compared with main effect analysis, interaction analysis also needs a larger sample size. Thus, little attention has been paid to interaction analysis for rare genetic variants, possibly due to the limited sample size in many cohort studies.

In this chapter, we propose a general approach for testing gene by environment interaction on quantitative traits in a SKAT framework. We start from 3 ways to obtain residuals from the null model, and show that 2 of them are equivalent. We compare our two approaches with a burden test of the interaction term in simulation studies, and we

also illustrate our approaches in testing gene by BMI interaction on fasting glucose, adjusting for age and sex, using an unrelated subset of individuals from the Framingham Heart Study.

4.2 Burden Test for Gene-Environment Interaction

Assuming a sample size of n , let the $n \times 1$ vector of the quantitative trait \mathbf{y} follow a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \mathbf{E}\mathbf{G}\mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where \mathbf{X} is an $n \times p$ covariate matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector consisting of fixed effects parameters for covariates (an intercept and $p - 1$ coefficients for covariates), \mathbf{G} is an $n \times q$ genotype matrix for q rare genetic variants of interest, \mathbf{W} is a $q \times q$ pre-specified diagonal weight matrix for the rare variants, $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the main effect of rare variants, \mathbf{E} is an $n \times n$ diagonal matrix with elements the environmental variable of interest, which is included in \mathbf{X} as a column, with mean 0, $\boldsymbol{\delta}$ is a $q \times 1$ vector for the SNP by environment interaction effect, and it is equal to $\delta\mathbf{1}$, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the error. The vector of error $\boldsymbol{\varepsilon}$ follows

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n),$$

where σ_E^2 is the variance parameter. In this parameter setting, we are interested in testing $H_0: \boldsymbol{\delta} = \mathbf{0}$ versus $H_1: \boldsymbol{\delta} \neq \mathbf{0}$. This is a burden test, because we implicitly assume that all rare variants in this test share the same interaction effect size (after weighting).

4.3 Sequence Kernel Association Test for Gene-Environment Interaction

We follow the same notations as in Section 4.2, but now $\boldsymbol{\delta}$ is a $q \times 1$ vector for the random effects of gene-environment interaction. The vector of error $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\delta}$ are assumed normally distributed and uncorrelated with each other:

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_I^2 \mathbf{I}_q),$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n).$$

where σ_I^2 and σ_E^2 are corresponding variance component parameters. We are interested in testing $H_0: \sigma_I^2 = 0$ versus $H_1: \sigma_I^2 > 0$, which is equivalent to testing $H_0: \boldsymbol{\delta} = \mathbf{0}$ versus $H_1: \boldsymbol{\delta} \neq \mathbf{0}$. Here we propose 3 ways to obtain the residuals from the null model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

4.3.1 Fixed Main Effects

For the first approach to obtain residuals, we treat the genotype main effects $\boldsymbol{\gamma}$ as fixed effects, and fit the null model as a linear regression model with covariates and genotype main effects. We fit the null linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

and obtain estimates $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\sigma}_E^2$. Under the null hypothesis of no interaction, the test statistic

$$Q_{FIX} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{G}\mathbf{W}\widehat{\boldsymbol{\gamma}})' \mathbf{E}\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}'\mathbf{E}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{G}\mathbf{W}\widehat{\boldsymbol{\gamma}}),$$

follows a weighted sum of χ_1^2 distribution

$$Q_{FIX} \sim \sum_{i=1}^q \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix

$$\Psi_{FIX} = \hat{\sigma}_E^2 \mathbf{W} \mathbf{G}' \mathbf{E} (\mathbf{I}_n - \tilde{\mathbf{X}} (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}') \mathbf{E} \mathbf{G} \mathbf{W},$$

where $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{G} \mathbf{W})_{n \times (p+q)}$ is the combined matrix for covariates and weighted genotypes. In this chapter, we denote this approach as SKAT-FIX.

4.3.2 Random Main Effects with Residuals Adjusting for Covariates Only

In our second approach, we assume a linear mixed effects model

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{G} \mathbf{W} \boldsymbol{\gamma} + \mathbf{E} \mathbf{G} \mathbf{W} \boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where random genotype main effects $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_M^2 \mathbf{I}_q)$. We also assume that $\boldsymbol{\gamma}$ is

uncorrelated with random effects for the interaction $\boldsymbol{\delta}$ and error $\boldsymbol{\varepsilon}$. We obtain estimates $\hat{\boldsymbol{\beta}}$,

$\hat{\sigma}_M^2$, $\hat{\sigma}_E^2$ by fitting the null linear mixed effects model

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{G} \mathbf{W} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Under the null hypothesis of no interaction, the test statistic

$$Q_{RAN} = (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{E} \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' \mathbf{E} \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}),$$

follows a weighted sum of χ_1^2 distribution

$$Q_{RAN} \sim \sum_{i=1}^q \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix

$$\Psi_{RAN} = \mathbf{W} \mathbf{G}' \mathbf{E} (\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X} (\mathbf{X}' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\boldsymbol{\Sigma}}^{-1}) \mathbf{E} \mathbf{G} \mathbf{W},$$

where $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_M^2 \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' + \hat{\sigma}_E^2 \mathbf{I}_n$. In this chapter, we denote this approach as SKAT-RAN.

4.3.3 Random Main Effects with Residuals Adjusting for Genotype Main Effects

For the third approach, we fit the same null model as in Section 4.3.2, but instead of using residuals $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, we now use residuals $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}}$, where $\hat{\boldsymbol{\gamma}}$ is estimated random effects for the genotype main effects. The difference is that in Section 4.3.2, we consider the genotype main effects in the matrix $\hat{\boldsymbol{\Sigma}}$ through the parameter estimate $\hat{\sigma}_M^2$. Now we have adjusted the genotype main effects in the residuals, the test statistic becomes

$$Q_{RAN'} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}})' \mathbf{E} \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' \mathbf{E} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}}).$$

Given the fact that

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}, \\ \hat{\boldsymbol{\gamma}} &= (\hat{\sigma}_M^2\mathbf{I}_q)\mathbf{W}\mathbf{G}'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{aligned}$$

we have

$$\begin{aligned} \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\sigma}_M^2\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\hat{\boldsymbol{\Sigma}} - \hat{\sigma}_M^2\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}')\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\hat{\sigma}_E^2\mathbf{I}_n)\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{aligned}$$

thus

$$\begin{aligned} Q_{RAN'} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}})' \mathbf{E} \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' \mathbf{E} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}}) \\ &= \hat{\sigma}_E^4 (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{E} \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}' \mathbf{E} \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \hat{\sigma}_E^4 Q_{RAN}. \end{aligned}$$

This test statistic is proportional to SKAT-RAN test statistic we derive in Section 4.3.2. In this chapter, we do not consider it as a different approach, and we do not include it in simulation studies or real data analysis.

4.4 Extension to Related Individuals

The methods described in Sections 4.2 and 4.3 are only applicable to unrelated individuals. However, some cohort studies include related individuals. In this section, we discuss the extension of burden test and SKAT in testing gene-environment interaction to related individuals with known family structure.

4.4.1 Burden Test

In samples with related individuals, our revised model is a linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \mathbf{E}\mathbf{G}\mathbf{W}\boldsymbol{\delta} + \boldsymbol{\omega} + \boldsymbol{\varepsilon},$$

where \mathbf{y} is an $n \times 1$ vector of the quantitative trait, \mathbf{X} is an $n \times p$ covariate matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector consisting of fixed effects parameters for covariates (an intercept and $p - 1$ coefficients for covariates), \mathbf{G} is an $n \times q$ genotype matrix for q rare genetic variants of interest, \mathbf{W} is a $q \times q$ pre-specified diagonal weight matrix for the rare variants, $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the main effect of rare variants, \mathbf{E} is an $n \times n$ diagonal matrix with the environmental variable of interest, which has mean 0 and is included in \mathbf{X} as a column, $\boldsymbol{\delta}$ is a $q \times 1$ vector for the SNP by environment interaction effect, and it is equal to $\boldsymbol{\delta}\mathbf{1}$, $\boldsymbol{\omega}$ is an $n \times 1$ vector for the random effects of familial correlation, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for

the error. The vector of error $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\omega}$ are assumed normally distributed and uncorrelated with each other:

$$\begin{aligned}\boldsymbol{\omega} &\sim N(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi}), \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n),\end{aligned}$$

where $\boldsymbol{\Phi}$ is twice the kinship matrix of size $n \times n$ obtained from family information only, σ_G^2, σ_E^2 are corresponding variance component parameters. In this parameter setting, we are interested in testing $H_0: \delta = 0$ versus $H_1: \delta \neq 0$. This is a test for a fixed effect parameter in a linear mixed effects model.

4.4.2 SKAT with Fixed Main Effects

We follow the same notations as in Section 4.4.1, but now $\boldsymbol{\delta}$ is a $q \times 1$ vector for the random effects of gene-environment interaction. The vector of error $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ are assumed normally distributed and uncorrelated with each other:

$$\begin{aligned}\boldsymbol{\delta} &\sim N(\mathbf{0}, \sigma_I^2 \mathbf{I}_q), \\ \boldsymbol{\omega} &\sim N(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi}), \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n).\end{aligned}$$

where σ_I^2, σ_G^2 and σ_E^2 are corresponding variance component parameters. We are interested in testing $H_0: \sigma_I^2 = 0$ versus $H_1: \sigma_I^2 > 0$, which is equivalent to testing $H_0: \boldsymbol{\delta} = \mathbf{0}$ versus $H_1: \boldsymbol{\delta} \neq \mathbf{0}$. We first fit the null linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\omega} + \boldsymbol{\varepsilon},$$

and obtain estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_G^2, \hat{\sigma}_E^2$. We define

$$\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_G^2 \boldsymbol{\Phi} + \hat{\sigma}_E^2 \mathbf{I}_n.$$

Under the null hypothesis of no interaction, the test statistic

$$Q_{FIX} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}})' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{E} \mathbf{G} \mathbf{W} \mathbf{W}' \mathbf{G}' \mathbf{E}' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{GW}\hat{\boldsymbol{\gamma}}),$$

follows a weighted sum of χ_1^2 distribution

$$Q_{FIX} \sim \sum_{i=1}^q \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix

$$\boldsymbol{\Psi}_{FIX} = \mathbf{W} \mathbf{G}' \mathbf{E} (\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{X}} (\tilde{\mathbf{X}}' \hat{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \hat{\boldsymbol{\Sigma}}^{-1}) \mathbf{E} \mathbf{G} \mathbf{W},$$

where $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{GW})_{n \times (p+q)}$ is the combined matrix for covariates and weighted genotypes.

4.4.3 SKAT with Random Main Effects

We follow the same notations as in Section 4.4.2, and in addition, we assume $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of genotype main effects. The vector of error $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ are assumed normally distributed and uncorrelated with each other:

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_M^2 \mathbf{I}_q),$$

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_I^2 \mathbf{I}_q),$$

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi}),$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_E^2 \mathbf{I}_n).$$

where σ_M^2 , σ_I^2 , σ_G^2 and σ_E^2 are corresponding variance component parameters. We are interested in testing $H_0: \sigma_I^2 = 0$ versus $H_1: \sigma_I^2 > 0$, which is equivalent to testing $H_0: \boldsymbol{\delta} = \mathbf{0}$ versus $H_1: \boldsymbol{\delta} \neq \mathbf{0}$. We first fit the null linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\omega} + \boldsymbol{\varepsilon},$$

and obtain estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_M^2$, $\widehat{\sigma}_G^2$, $\widehat{\sigma}_E^2$. We define

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\sigma}_M^2 \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}' + \widehat{\sigma}_G^2 \boldsymbol{\Phi} + \widehat{\sigma}_E^2 \mathbf{I}_n.$$

Under the null hypothesis of no interaction, the test statistic

$$Q_{RAN} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{E}\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}' \mathbf{E} \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

follows a weighted sum of χ_1^2 distribution

$$Q_{RAN} \sim \sum_{i=1}^q \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix

$$\boldsymbol{\Psi}_{RAN} = \mathbf{W}\mathbf{G}' \mathbf{E} \left(\widehat{\boldsymbol{\Sigma}}^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{X} (\mathbf{X}' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \widehat{\boldsymbol{\Sigma}}^{-1} \right) \mathbf{E}\mathbf{G}\mathbf{W}.$$

It is easy to show that for related individuals, if we adjust for the genotype main effects and use a Q_{FIX} type of test statistic, as we describe for unrelated individuals in Section 4.3.3, we would get exactly the same test statistic as Q_{RAN} .

4.5 Simulation Studies

4.5.1 Type I Error

4.5.1.1 Simulation Design

To evaluate the performance of the burden test and two SKAT approaches in detecting gene-environment interaction, we first performed two null simulation studies: 1. The trait is not associated with the genotypes; 2. The trait is associated with the genotypes but there is no interaction effects. For each scenario, we simulated 100 genotype datasets with 2000 unrelated individuals and 20 single nucleotide polymorphisms (SNP) with minor allele frequency (MAF) randomly sampled from a uniform distribution of 0.005 to 0.05, and with low ($r = 0.1$), moderate ($r = 0.5$), high ($r = 0.7$) linkage disequilibrium (LD) between adjacent SNPs. The correlation between farther SNPs decays as an autoregressive model with order 1. In the first null simulation study, for each genotype dataset, 1000 phenotype datasets including covariates were simulated from

$$y = 0.05\mathbf{age} + 0.5\mathbf{sex} + 0.1\mathbf{bmi} + \boldsymbol{\varepsilon},$$

where \mathbf{age} is a vector of continuous covariate generated from a normal distribution with mean 50 and standard deviation 5, \mathbf{sex} is a vector of dichotomous covariate generated from a Bernoulli distribution with probability 0.5, \mathbf{bmi} is a vector of continuous covariate generated from a normal distribution with mean 25 and standard deviation 4, $\boldsymbol{\varepsilon}$ is independent and identically distributed standard normal random error.

In the second null simulation study, genotype datasets were generated the same way as in the first study, but we fixed LD between adjacent SNPs to be $r = 0.5$. For each genotype dataset, 1000 phenotype datasets were simulated from

$$\mathbf{y} = 0.05\mathbf{age} + 0.5\mathbf{sex} + 0.1\mathbf{bmi} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where \mathbf{age} , \mathbf{sex} , \mathbf{bmi} and $\boldsymbol{\varepsilon}$ were generated in the same way as before, $\boldsymbol{\gamma}$ is a vector consisting of the effect sizes of the causal SNPs. We varied the proportion of causal SNPs from 20% to 50% and 80%, and we simulated both same and opposite directions of main effects. Causal SNPs were randomly selected out of the 20 SNPs for each phenotype replicate, and in each parameter setting the effect sizes of causal SNPs were determined by

$$\gamma_i = \sqrt{\frac{c}{2MAF_i(1 - MAF_i)}},$$

where MAF_i is the MAF used to generate the genotype dataset for causal SNP i , and c is a constant for all causal SNPs in each phenotype replicate, calculated as

$$c = \frac{R^2}{\mathbf{v}'\mathbf{D}\mathbf{v}},$$

where R^2 , the total proportion of variance explained by all causal SNPs, was fixed at 1% for scenarios when all causal SNPs had main effects in the same direction, and 5% for scenarios when 50% of the causal SNPs had positive main effects and 50% had negative main effects. \mathbf{D} is the LD correlation matrix for the 20 SNPs, and \mathbf{v} is a vector indicating the directions of causal SNP effects in each replicate.

In both null simulation studies, we tested gene by BMI interaction. We used Wu weights for all three approaches.

4.5.1.2 Simulation Results

Table 4.1 shows the empirical type I errors from 100,000 replicates in the first null simulation study without genotype main effects, using the burden test (BT), SKAT-FIX and SKAT-RAN for gene by BMI interaction. At the α levels of 0.05, 0.01 and 0.001, all three methods have correct type I errors. The conclusion is consistent regardless of the LD structure in the genotype data. The Quantile-Quantile plots from the first null simulation study without genotype main effects are shown in Figure 4.1. In all 3 scenarios, the p-values from BT, SKAT-FIX and SKAT-RAN are all very close to the expected uniform distribution on the interval (0, 1).

Table 4.2 shows the empirical type I errors from 100,000 replicates in the second null simulation study with genotype main effects, using BT, SKAT-FIX and SKAT-RAN for gene by BMI interaction. At all three α levels of 0.05, 0.01 and 0.001, we did not observe strong evidence for inflated type I errors, for any of the three methods. The results suggest that the conclusion does not depend on the proportion of causal markers with main effects, or the proportion of protective and detrimental genetic markers. Figure 4.2 shows the Quantile-Quantile plots from the second null simulation with genotype main effects in the same direction, and Figure 4.3 shows the Quantile-Quantile plots from the second null simulation with genotype main effects in different directions. In all 6

scenarios, the p-values from BT, SKAT-FIX and SKAT-RAN are all very close to the expected uniform distribution on the interval (0, 1).

Results from both simulation studies suggest that all three methods are valid in type I errors in various scenarios, with or without genotype main effects.

Table 4.1 Type I errors from the null simulation without genotype main effects

LD	α level	BT	SKAT-FIX	SKAT-RAN
r = 0.1	0.05	0.0503	0.0497	0.0496
	0.01	0.0097	0.0100	0.0098
	0.001	0.0009	0.0010	0.0010
r = 0.5	0.05	0.0498	0.0509	0.0498
	0.01	0.0097	0.0102	0.0101
	0.001	0.0010	0.0011	0.0010
r = 0.7	0.05	0.0506	0.0519	0.0505
	0.01	0.0100	0.0097	0.0096
	0.001	0.0010	0.0009	0.0008

Empirical type I errors were calculated as the proportion of p-values less than or equal to the corresponding α level in 100,000 genotype-phenotype datasets.

Table 4.2 Type I errors from the null simulation with genotype main effects

Causal markers	α level	BT	SKAT-FIX	SKAT-RAN
+/-/0 = 4/0/16	0.05	0.0502	0.0507	0.0500
	0.01	0.0100	0.0096	0.0096
	0.001	0.0007	0.0009	0.0009
+/-/0 = 10/0/10	0.05	0.0502	0.0508	0.0503
	0.01	0.0103	0.0100	0.0099
	0.001	0.0010	0.0011	0.0011
+/-/0 = 16/0/4	0.05	0.0501	0.0513	0.0509
	0.01	0.0098	0.0104	0.0099
	0.001	0.0010	0.0008	0.0008
+/-/0 = 2/2/16	0.05	0.0497	0.0501	0.0506
	0.01	0.0097	0.0098	0.0098
	0.001	0.0009	0.0010	0.0010
+/-/0 = 5/5/10	0.05	0.0496	0.0512	0.0511
	0.01	0.0097	0.0104	0.0105
	0.001	0.0010	0.0008	0.0009
+/-/0 = 8/8/4	0.05	0.0511	0.0516	0.0518
	0.01	0.0105	0.0103	0.0099
	0.001	0.0009	0.0009	0.0009

Empirical type I errors were calculated as the proportion of p-values less than or equal to the corresponding α level in 100,000 genotype-phenotype datasets. +/-/0 denote the number of causal markers with positive and negative effects, and neutral markers.

Figure 4.1 Quantile-Quantile plots from the null simulation without genotype main effects

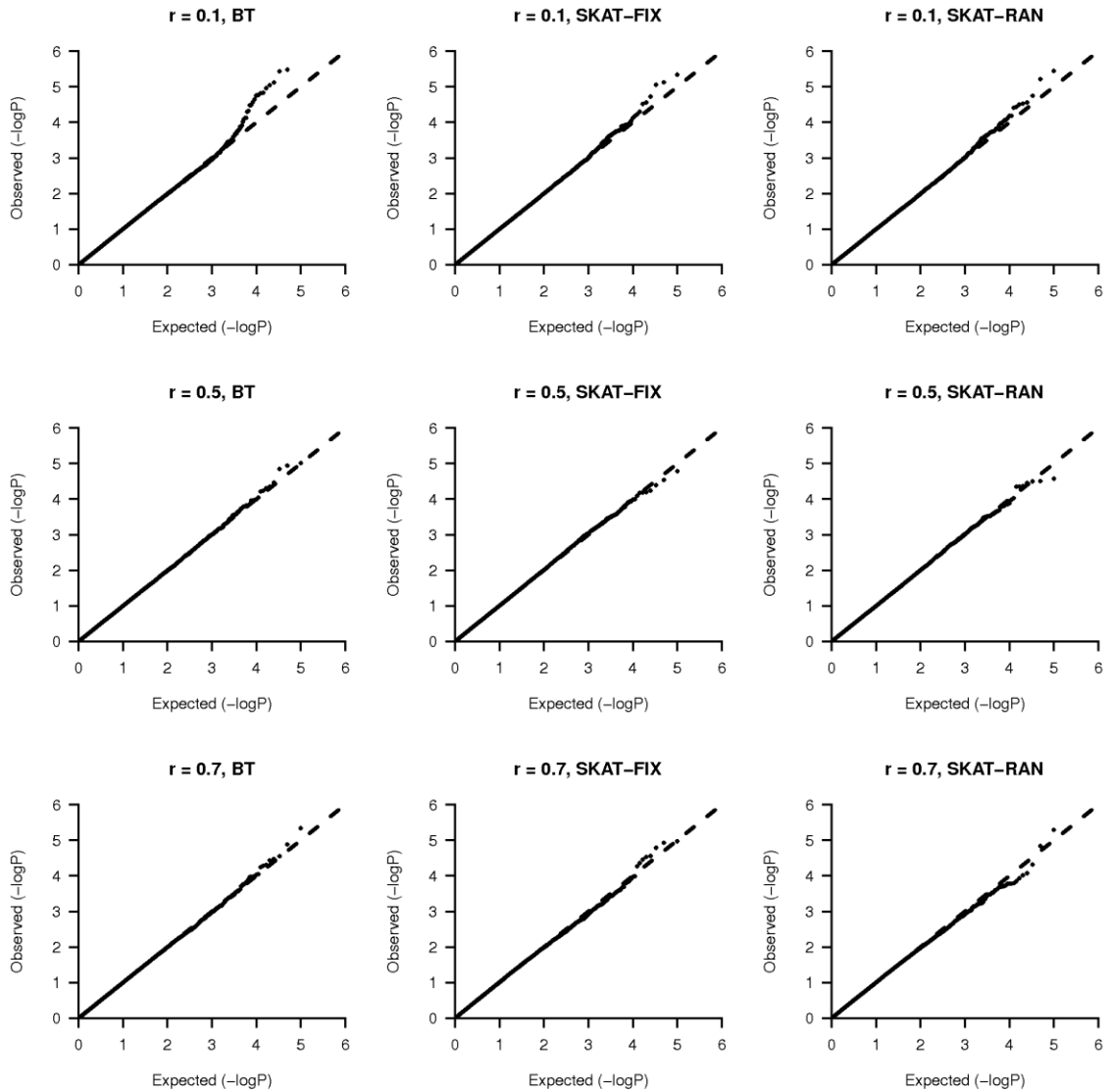


Figure 4.2 Quantile-Quantile plots from the null simulation with genotype main effects in the same direction

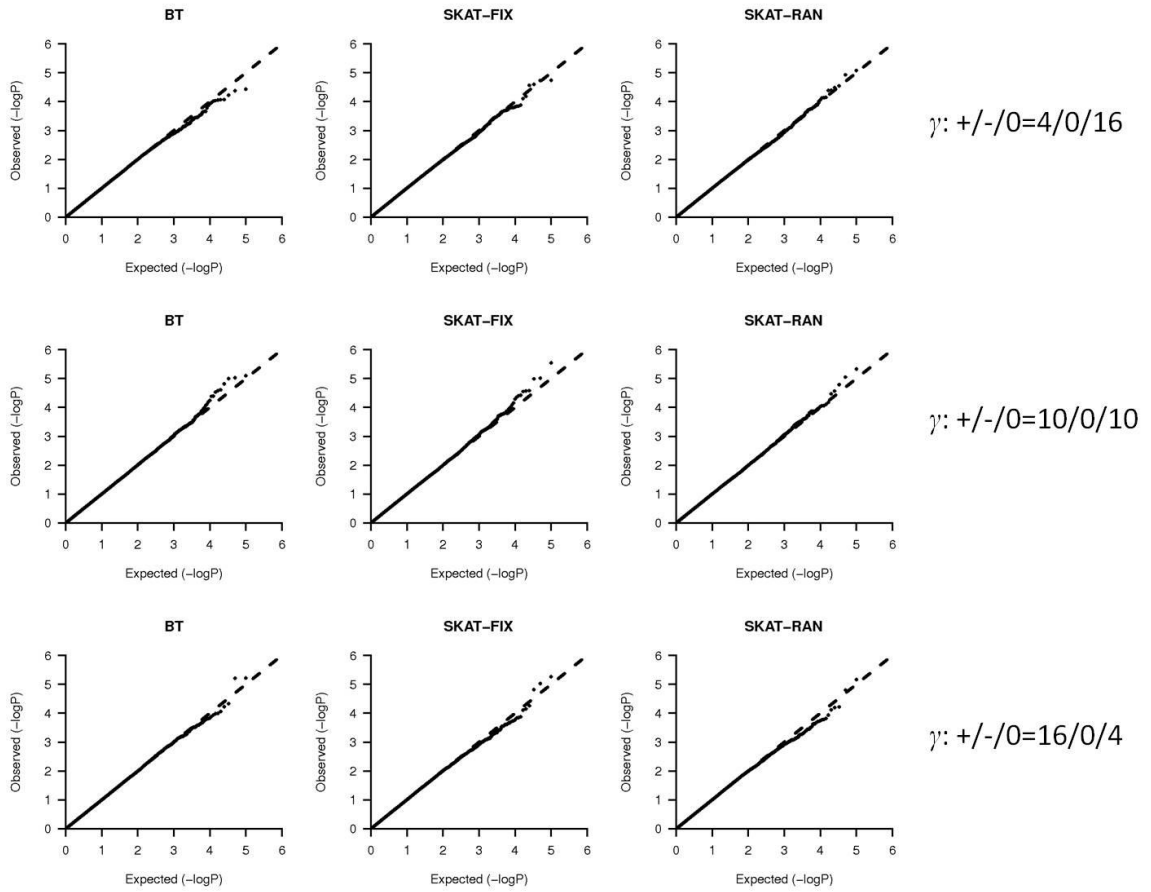
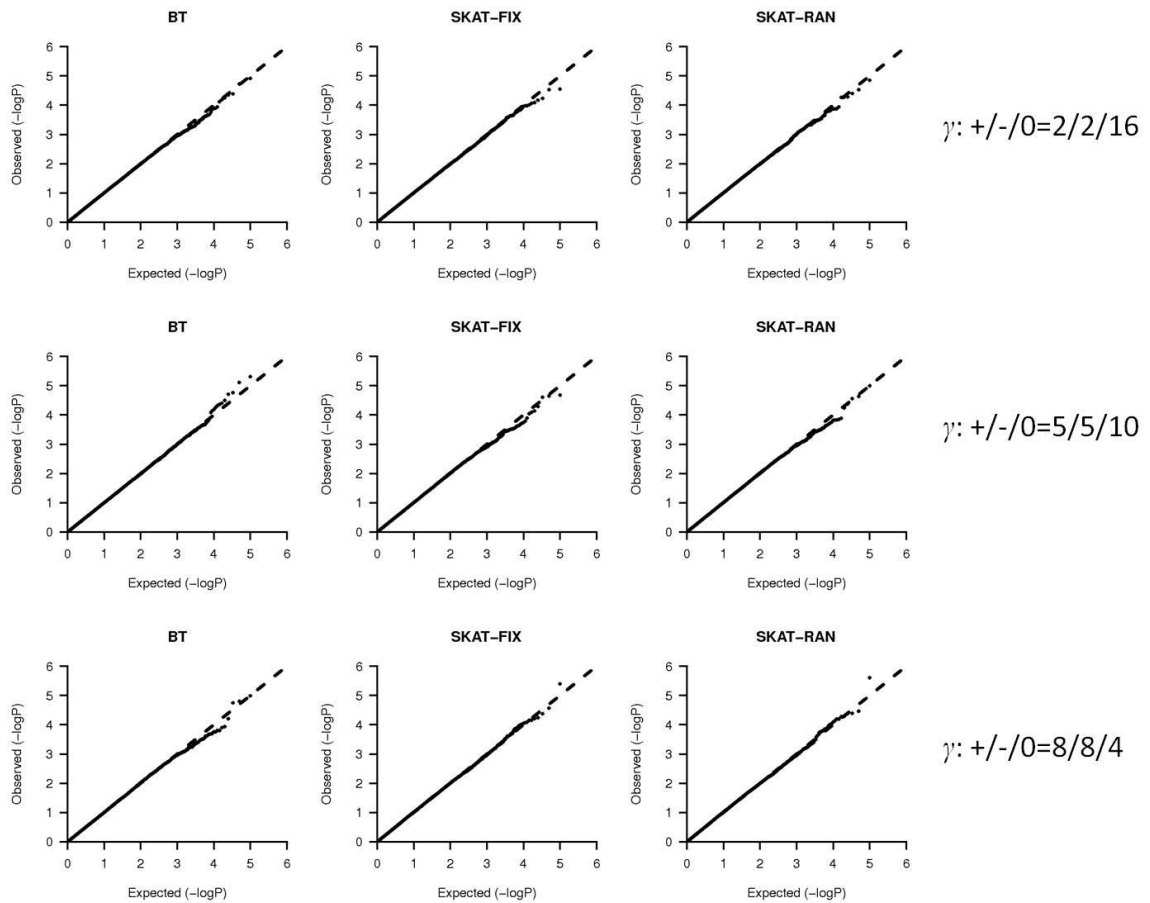


Figure 4.3 Quantile-Quantile plots from the null simulation with genotype main effects in different directions



4.5.2 Power

4.5.2.1 Simulation Design

To evaluate the power of the burden test and two SKAT approaches in detecting gene-environment interaction, we performed two simulation studies under the alternative hypothesis: 1. There is gene by BMI interaction; 2. There is gene by sex interaction. We investigated the performance of the three approaches in detecting gene by environmental

variable interaction using both continuous and dichotomous environmental variables. For each scenario, we simulated 100 genotype datasets with 2000 unrelated individuals and 20 SNPs with MAF randomly sampled from a uniform distribution of 0.005 to 0.05, and we fixed LD between adjacent SNPs to be $r = 0.5$.

In the first simulation study of gene by BMI interaction, for each genotype dataset, 1000 phenotype datasets were simulated from

$$\mathbf{y} = 0.05\mathbf{age} + 0.5\mathbf{sex} + 0.1\mathbf{bmi} + \mathbf{G}\boldsymbol{\gamma} + (\mathbf{bmi} - 25)\mathbf{G}\boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where \mathbf{age} , \mathbf{sex} , \mathbf{bmi} and $\boldsymbol{\varepsilon}$ were generated in the same way as in the null simulation studies, $\boldsymbol{\gamma}$ was determined using the same method as in the second null simulation study with genotype main effects. We varied the proportion of causal SNPs from 20% to 50% and 80%, and we simulated both same and opposite directions of main effects. We also simulated both same and opposite directions of interaction effects. Causal SNPs were randomly selected out of the 20 SNPs for each phenotype replicate, and in each parameter setting the interaction effect sizes of causal SNPs were determined by

$$\delta_i = \sqrt{\frac{c}{2MAF_i(1 - MAF_i)Var(bmi)'}}$$

where MAF_i is the MAF used to generate the genotype dataset for causal SNP i , and c is a constant for all causal SNPs in each phenotype replicate, calculated as

$$c = \frac{R^2}{\mathbf{v}'\mathbf{D}\mathbf{v}'},$$

where R^2 , the total proportion of variance explained by gene by BMI interaction, was fixed at 1% for scenarios when all causal SNPs had interaction effects in the same direction, and 5% for scenarios when 50% of the causal SNPs had positive interaction effects and 50% had negative interaction effects. \mathbf{D} is the LD correlation matrix for the 20 SNPs, and \mathbf{v} is a vector indicating the directions of causal SNP interaction effects in each replicate. We tested gene by BMI interaction and used Wu weights for all three approaches.

In the second simulation study of gene by sex interaction, for each genotype dataset, 1000 phenotype datasets were simulated from

$$\mathbf{y} = 0.05\mathbf{age} + 0.5\mathbf{sex} + 0.1\mathbf{bmi} + \mathbf{G}\boldsymbol{\gamma} + (\mathbf{sex} - 0.5)\mathbf{G}\boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where \mathbf{age} , \mathbf{sex} , \mathbf{bmi} , $\boldsymbol{\varepsilon}$ and $\boldsymbol{\gamma}$ were generated in the same way as in the first simulation study of gene by BMI interaction. We varied the proportion of causal SNPs from 20% to 50% and 80%, and we simulated both same and opposite directions of main effects. We also simulated both same and opposite directions of interaction effects. Causal SNPs were randomly selected out of the 20 SNPs for each phenotype replicate, and in each parameter setting the interaction effect sizes of causal SNPs were determined by

$$\delta_i = \sqrt{\frac{c}{2MAF_i(1 - MAF_i)Var(\mathbf{sex})}},$$

where MAF_i is the MAF used to generate the genotype dataset for causal SNP i , and c is a constant for all causal SNPs in each phenotype replicate, calculated as

$$c = \frac{R^2}{\mathbf{v}'\mathbf{D}\mathbf{v}},$$

where R^2 , \mathbf{D} and \mathbf{v} were determined in the same way as in the first simulation study of gene by BMI interaction. We tested gene by sex interaction and used Wu weights for all three approaches.

4.5.2.2 Simulation Results

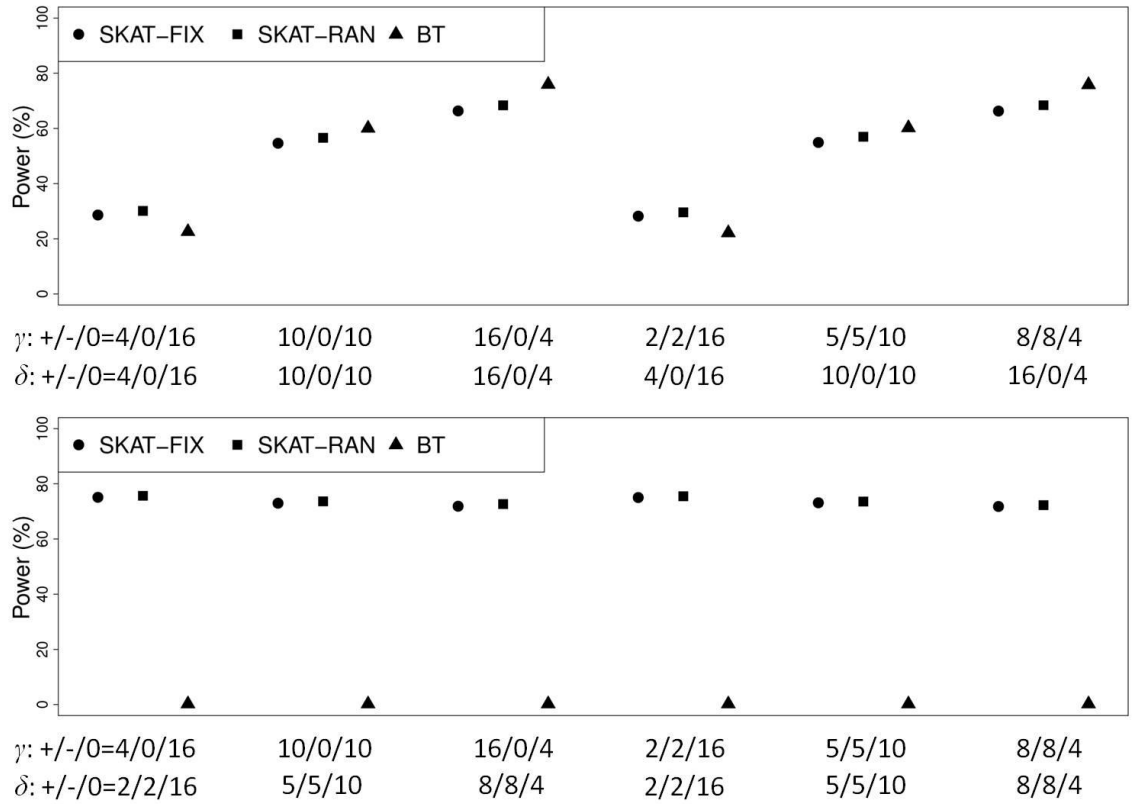
Figure 4.4 shows the power results from the first simulation study of gene by BMI interaction. Empirical power was calculated at the significance level of 10^{-5} . Figure 4.4 suggests that the direction of main effects $\boldsymbol{\gamma}$ almost has no effect on power. When the proportions of causal markers with positive interaction effects, negative interaction effects, and neutral markers are fixed, the power remains almost the same no matter whether the causal markers have the same or different directions of main effects. The burden test has the highest power when the proportion of causal markers is large and all causal markers have interaction effects in the same direction. However, it has almost no power when causal markers have interaction effects in different directions, which matches our expectation. The SKAT-type tests are most powerful when the proportion of causal markers is small, or when causal markers have interaction effects in different directions. SKAT-RAN has slightly higher power than SKAT-FIX in all scenarios, but the difference in power is trivial.

Figure 4.5 shows the power results from the second simulation study of gene by sex interaction, at the significance level of 10^{-5} . The results are consistent with Figure 4.4,

suggesting that each of these methods performs similarly when analyzing gene by continuous covariate interaction and gene by dichotomous covariate interaction.

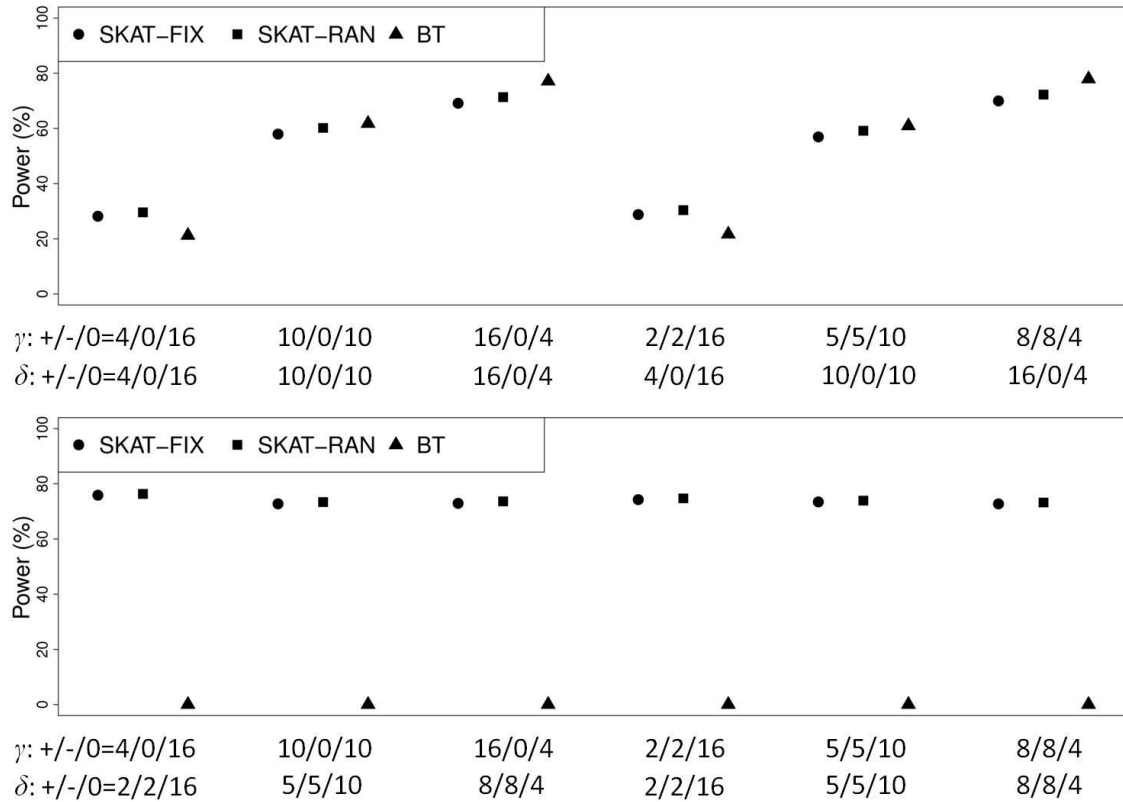
Generally, the SKAT-type tests perform well in all the scenarios, and the performance of each test greatly depends on the proportions of positive, negative and neutral interaction effects δ , but not the proportions of positive, negative and neutral main effects γ .

Figure 4.4 Power comparisons of SKAT-FIX, SKAT-RAN and BT in detecting gene by BMI interaction



Empirical power calculated at α level of 10^{-5} . In each scenario, +/-0 indicates the number of SNPs with positive effects, negative effects and no effects. γ denotes SNP main effects, and δ denotes SNP by BMI interaction effects.

Figure 4.5 Power comparisons of SKAT-FIX, SKAT-RAN and BT in detecting gene by sex interaction



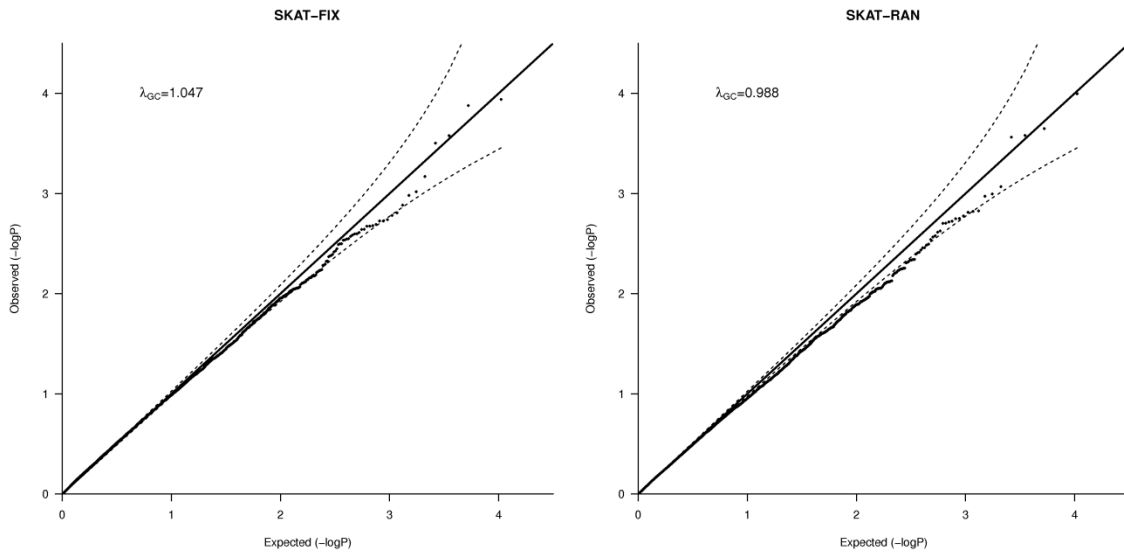
Empirical power calculated at α level of 10^{-5} . In each scenario, +/-/0 indicates the number of SNPs with positive effects, negative effects and no effects. γ denotes SNP main effects, and δ denotes SNP by sex interaction effects.

4.6 Application

We performed a genome-wide sliding window analysis to detect gene by BMI interaction on fasting glucose, using SHARe genotype data. We analyzed an unrelated subset of individuals, with total sample size 1921. We selected SNPs with MAF less than 5% and

ran the analysis using a sliding window of 500kb, with 250kb overlap each with previous and subsequent windows. We removed windows with 0 or 1 SNP, resulting in 10,538 windows for all autosomes with the number of SNPs ranging from 2 to 74 with median 18. No window reached the genome-wide significance using SKAT-FIX, SKAT-RAN or BT. The Quantile-Quantile plots for SKAT-FIX and SKAT-RAN are shown in Figure 4.6. There is no inflation of the p-values from this genome-wide analysis.

Figure 4.6 Quantile-Quantile plots for SKAT-type tests in the genome-wide sliding window analysis for gene by BMI interaction on fasting glucose



The p-values were plotted as minus log base 10 p-values. The genomic control factor λ_{GC} was computed as the ratio of median chi-square statistics with 1 df corresponding to observed and expected p-values.

4.7 Discussion

In this chapter, we propose two SKAT-type tests for detecting gene by environment interaction. Depending on whether to adjust for genotype main effects as fixed or random effects, the two tests are named SKAT-FIX and SKAT-RAN, respectively. We demonstrate that both of them are flexible and powerful tests on the gene by environment interaction effects, in the context of rare genetic variants analysis. When the number of rare variants in the test is large, we recommend SKAT-RAN over SKAT-FIX, not only because it has slightly higher power in general, but also because fitting a multiple linear regression model with a lot of predictors in SKAT-FIX may not be a good way to obtain residuals.

Compared with burden tests on the interaction effects, SKAT-type tests have higher power when the proportion of causal genetic markers in a genomic region is small, or when causal genetic markers have different directions of gene by environment interaction effects. Although the burden test performs better than SKAT-type tests when the proportion of causal genetic markers is greater than or equal to 50% and all these markers have the same direction of gene by environment interaction effects, SKAT-type tests do not suffer from a great power loss in these scenarios. In addition, external weights based on gene annotation and biological functional prediction can be incorporated to further increase the statistical power.

We did not find any gene by BMI interaction on fasting glucose in a genome-wide sliding window analysis using SHARe genotype data, at the genome-wide significance level. Since the SHARe project was not designed for rare variants analysis, most SNPs in our genotype dataset are common genetic markers with MAF greater than 5%, which were excluded from the analysis. With decreasing sequencing cost, we should be able to revisit our real data example in the future when more rare genetic variants are identified in the Framingham Heart Study.

In this chapter, we also derive the test statistics and null distributions of the two SKAT-type tests in analyzing related individuals, as an extension of famSKAT in Chapter 3 to tests for gene by environment interaction. When we have related individuals in our dataset, we do not need to exclude some of them to obtain an unrelated subset to perform the analysis. We have shown in Chapter 3 that selecting unrelated individuals suffers from power loss due to sample size reduction in the context of genotype main effects rare variants analysis, compared with famSKAT. We believe the situation is similar in gene by environment interaction tests. By using famSKAT-type tests, we do not have to reduce the sample size, and it is easy to see that famSKAT-type tests in Section 4.4 are the same as their corresponding SKAT-type equivalent in Section 4.3 when all individuals are unrelated.

We note that although the SKAT-type tests are generally powerful in many scenarios, they are not optimal in some cases. SKAT assumes that the genotype random effects of

rare variants have mean 0. When the true mean is not 0, the power may be compromised, compared with burden tests. Wang, Chen and Yang [2012] proposed different approaches, allowing for non-zero mean, and jointly tested the mean and the variance component. They showed in extensive simulation studies that their approaches outperform SKAT in many simulation scenarios. In the context of gene by environment interaction tests, we can also allow for non-zero means of interaction random effects $\boldsymbol{\delta}$, and jointly test the mean and variance component parameters. When adjusting for genotype main effects $\boldsymbol{\gamma}$ in SKAT-RAN, instead of $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_M^2 \mathbf{I}_q)$, we can also assume $\boldsymbol{\gamma} \sim N(\mu \mathbf{1}, \sigma_M^2 \mathbf{I}_q)$ and estimate μ from the null model. Lee, Wu and Lin [2012] dealt with the power loss of SKAT compared to burden tests in some scenarios and proposed a hybrid of SKAT and burden tests named SKAT-O. They showed that SKAT-O is better than SKAT when the proportion of causal genetic variants is large and all causal variants have the same direction of effects. We can use the same idea and derive SKAT-O-type tests for gene by environment interaction, which may improve the performance in some scenarios.

Chapter 5 Summary and Future Work

5.1 Summary

Genetic association studies have been of great interest in statistical genetics for years. However, for complex diseases and quantitative traits, all associated genetic variants identified so far only explain a small proportion of the heritability. We believe that the hunting for novel genetic variants accounting for the unexplained heritability is essential in fully explaining the disease etiology, and that novel statistical methods are strongly needed in this field.

In this dissertation, we propose general and specific statistical approaches which can be applied in genetic association studies. In Chapter 2, we develop a method of moments estimator for the between-study covariance matrix in random effects model multivariate meta-analysis. We hope this approach will facilitate random effects model multivariate meta-analysis in various scientific fields including statistical genetics and address the heterogeneity issue in the fixed effects model, described in Section 1.3. In Chapter 3, we extend SKAT to be applicable to rare genetic variants analysis on quantitative traits in family samples. We start from a general powerful and computationally efficient approach in rare genetic variants analysis for unrelated individuals, SKAT, introduced in Section 1.4, and modify it using the same strategy as for related individuals in single marker GWAS, linear mixed effects models, discussed in Section 1.2. In Chapter 4, we derive two interaction-only SKAT-type tests for both unrelated and related individuals, which

combines the ideas in Sections 1.4 and 1.5, and Section 1.2. Together, we hope these novel methods will advance genetic association studies (Section 1.1), and other scientific fields in which meta-analysis, correlated data analysis, sparse data analysis, or interaction analysis are applicable.

5.2 Future Work

5.2.1 Extension of the Method of Moments Estimator

Meta-regression is a regression based meta-analysis approach to investigate whether particular covariates explain any of the heterogeneity of effects between studies [Thompson and Higgins, 2002]. Similar to meta-analysis, meta-regression includes both fixed effects model and random effects model, and it can be either univariate or multivariate. Without loss of generality, we assume a random effects meta-regression model which takes a similar notation to the random effects meta-analysis model in Section 2.3:

$$\mathbf{b}_{kp \times 1} = \mathbf{X}_{kp \times pq} \boldsymbol{\gamma}_{pq \times 1} + \boldsymbol{\delta}_{kp \times 1} + \mathbf{e}_{kp \times 1},$$

where k is the number of studies, p is the number of effect sizes, q is the number of regression parameters (including the intercept) for each effect size, which satisfies $q < k$. We can see clearly that if $q = 1$, then the model only has an intercept for each effect size, and is equivalent to the meta-analysis model. In the meta-regression context, it is not difficult to develop a similar method of moments estimator for the between-study covariance matrix.

Ma and Mazumdar [2011] proposed a nonparametric and non-iterative method to estimate the between-study covariance matrix in random effects model multivariate meta-analysis, based on the theory of U-statistic. However, similar to the method proposed by Jackson, White and Thompson [2010], their approach also requires matching each element of the between-study covariance matrix separately, which may not be invariant to reparametrization of effect sizes. Thus, a matrix form estimator based on the theory of U-statistic, with invariance property to linear transformation, may be desirable in this field.

5.2.2 Sequence Kernel Association Test for Dichotomous Traits in Family Samples

The original SKAT approach for unrelated individuals is a powerful and flexible method which can be applied to both quantitative and dichotomous traits, although the distribution of the test statistic under the null hypothesis is different for quantitative and dichotomous phenotypes. Our extended approach to related individuals, however, currently only applies to quantitative traits. The major problem is that although we can model the familial correlation as a random effect in the linear mixed effects model when analyzing quantitative traits, the variance-covariance structure for dichotomous traits is different. The situation is complicated if we apply a generalized linear mixed effects model in the analysis and treat the familial correlation as a random effect, because the link function is no longer the identity link, and we are likely to mis-specify the covariance structure after transformation.

The issue is not specific to SKAT in family samples. It also exists for single marker tests on dichotomous traits. Usually we use the generalized estimating equations with independent variance structure to perform single marker tests. Using the same strategy, in SKAT we can also fit the null model without any genetic effects, but incorporating familial correlation. After deriving the covariance matrix for the residuals from the null model, we can develop a test statistic and its null distribution. However, the performance in terms of type I error and power needs further investigation.

5.2.3 Joint Test of Genetic Main Effects and Gene by Environment Interaction for Rare Genetic Variants

In Chapter 4 we develop two SKAT-type interaction-only tests for gene by environment interaction in rare genetic variants analysis. As discussed in Section 1.5, research questions addressed by interaction-only tests and joint tests are different. If we focus on testing whether there is association with any rare genetic variants, allowing for gene by environment interaction, then a joint test may be desirable. Following the same notation as in Section 4.3.2, we assume a linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{GW}\boldsymbol{\gamma} + \mathbf{EGW}\boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where random genotype main effects $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_M^2 \mathbf{I}_q)$, and gene by environment interaction random effects $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_I^2 \mathbf{I}_q)$. Although we can assume the same weights \mathbf{W} in genotype main effects and interaction effects, generally we cannot assume that the variance component parameters σ_M^2 and σ_I^2 are the same. In this model, a joint test would test the hypotheses $H_0: \sigma_M^2 = \sigma_I^2 = 0$ versus $H_1: \sigma_M^2 > 0$ or $\sigma_I^2 > 0$. We note that when

$\sigma_I^2 = 0$, the test $H_0: \sigma_M^2 = 0$ versus $H_1: \sigma_M^2 > 0$, and when $\sigma_M^2 = 0$, the test $H_0: \sigma_I^2 = 0$ versus $H_1: \sigma_I^2 > 0$ both follow a SKAT-type weighted sum of χ_1^2 distribution in curved parameter spaces. However, since the joint test lies in the parameter space $\{(\sigma_M^2, \sigma_I^2): \sigma_M^2 \geq 0, \sigma_I^2 \geq 0\}$, the test statistic and its null distribution need further investigation.

Appendices

Appendix A Derivation of the Method of Moments Estimator

Let

$$\Psi = \text{Cov}(\widehat{\boldsymbol{\beta}}_F) = (\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1} = \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1},$$

then

$$\widehat{\boldsymbol{\beta}}_F = (\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{b} = \Psi \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1}\mathbf{b}_j.$$

It follows that

$$\begin{aligned} & \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1}(\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)(\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)' \\ &= \sum_{j=1}^k \left\{ \boldsymbol{\Sigma}_j^{-1}[(\mathbf{b}_j - \boldsymbol{\beta}) - (\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})][(\mathbf{b}_j - \boldsymbol{\beta}) - (\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})]' \right\} \\ &= \sum_{j=1}^k \left\{ \boldsymbol{\Sigma}_j^{-1} [(\mathbf{b}_j - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' - (\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' - (\mathbf{b}_j - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})' \right. \\ & \quad \left. + (\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})'] \right\}. \end{aligned}$$

Under the heterogeneity hypothesis

$$E(\mathbf{b}_j - \boldsymbol{\beta}) = 0,$$

$$E(\mathbf{b}_j - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' = \text{Var}(\mathbf{b}_j - \boldsymbol{\beta}) = \boldsymbol{\Sigma}_j + \mathbf{T}.$$

Assuming that any two studies are independent, then for $i \neq j$

$$E(\mathbf{b}_i - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' = 0.$$

So we can write

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' &= E\left(\boldsymbol{\Psi} \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} \mathbf{b}_i - \boldsymbol{\beta}\right)(\mathbf{b}_j - \boldsymbol{\beta})' \\ &= E\left[\boldsymbol{\Psi} \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} (\mathbf{b}_i - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})'\right] \\ &= E\left[\boldsymbol{\Psi} \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})'\right] \\ &= \boldsymbol{\Psi} + \boldsymbol{\Psi} \boldsymbol{\Sigma}_j^{-1} \mathbf{T}, \end{aligned}$$

$$E(\mathbf{b}_j - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})' = \left[E(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})'\right]' = \boldsymbol{\Psi} + \mathbf{T} \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Psi},$$

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})' &= E\left[\boldsymbol{\Psi} \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} (\mathbf{b}_i - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Psi}\right] \\ &= \boldsymbol{\Psi} \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\Sigma}_i + \mathbf{T}) \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Psi} \\ &= \boldsymbol{\Psi} \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Psi} + \boldsymbol{\Psi} \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} \mathbf{T} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Psi} = \boldsymbol{\Psi} + \boldsymbol{\Psi} \left(\sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1} \mathbf{T} \boldsymbol{\Sigma}_i^{-1}\right) \boldsymbol{\Psi}. \end{aligned}$$

It follows that

$$\begin{aligned}
& E(\mathbf{b}_j - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' - E(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\mathbf{b}_j - \boldsymbol{\beta})' - E(\mathbf{b}_j - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})' \\
& \quad + E(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})' \\
& = \boldsymbol{\Sigma}_j + \mathbf{T} - \boldsymbol{\Psi} - \boldsymbol{\Psi}\boldsymbol{\Sigma}_j^{-1}\mathbf{T} - \mathbf{T}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Psi} + \boldsymbol{\Psi}\left(\sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\mathbf{T}\boldsymbol{\Sigma}_i^{-1}\right)\boldsymbol{\Psi}, \\
& \quad E\left[\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1}(\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)(\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)'\right] \\
& = \sum_{j=1}^k \left[\mathbf{I}_{p \times p} + \boldsymbol{\Sigma}_j^{-1}\mathbf{T} - \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Psi} - \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}_j^{-1}\mathbf{T} - \boldsymbol{\Sigma}_j^{-1}\mathbf{T}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Psi} + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Psi}\left(\sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\mathbf{T}\boldsymbol{\Sigma}_i^{-1}\right)\boldsymbol{\Psi} \right] \\
& = k\mathbf{I}_{p \times p} + \boldsymbol{\Psi}^{-1}\mathbf{T} - \mathbf{I}_{p \times p} - \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}_i^{-1}\mathbf{T} - \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\mathbf{T}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Psi} \\
& \quad + \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1}\right)\boldsymbol{\Psi}\left(\sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\mathbf{T}\boldsymbol{\Sigma}_i^{-1}\right)\boldsymbol{\Psi} \\
& = (k-1)\mathbf{I}_{p \times p} + \left(\boldsymbol{\Psi}^{-1} - \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}_i^{-1}\right)\mathbf{T}.
\end{aligned}$$

Let

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}^{-1} - \sum_{i=1}^k \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}_i^{-1} = \sum_{i=1}^k \left[\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_i^{-1}\left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1}\right)^{-1}\boldsymbol{\Sigma}_i^{-1} \right],$$

then

$$E \left[\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F) (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)' - (k-1) \mathbf{I}_{p \times p} \right] = \boldsymbol{\Phi} \mathbf{T}.$$

Let

$$\mathbf{A} = \sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F) (\mathbf{b}_j - \widehat{\boldsymbol{\beta}}_F)' - (k-1) \mathbf{I}_{p \times p},$$

then

$$E(\boldsymbol{\Phi}^{-1} \mathbf{A}) = \mathbf{T}.$$

If we transpose it, we also have

$$E(\mathbf{A}' \boldsymbol{\Phi}^{-1}) = \mathbf{T},$$

then

$$E \left(\frac{\boldsymbol{\Phi}^{-1} \mathbf{A} + \mathbf{A}' \boldsymbol{\Phi}^{-1}}{2} \right) = \mathbf{T}.$$

$\boldsymbol{\Phi}$ is symmetric. Throughout this dissertation, we assume that the covariance matrix $\boldsymbol{\Sigma}_j$ is positive definite for all j . Otherwise if one or more covariance matrices have at least one eigenvalue of 0, the determinant of $\boldsymbol{\Sigma}$ would be 0, making it not invertible.

When all $\boldsymbol{\Sigma}_j$ are positive definite, all $\boldsymbol{\Sigma}_j^{-1}$ are also positive definite, $\boldsymbol{\Sigma}_j^{-1} \succ \mathbf{0}$, so

$$\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \succ \boldsymbol{\Sigma}_i^{-1},$$

$$\boldsymbol{\Sigma}_i = (\boldsymbol{\Sigma}_i^{-1})^{-1} \succ \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1},$$

$$\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_i^{-1} \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1} \boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_i^{-1} \left[\boldsymbol{\Sigma}_i - \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1} \right] \boldsymbol{\Sigma}_i^{-1} \succ 0,$$

$$\boldsymbol{\Phi} = \sum_{i=1}^k \left[\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_i^{-1} \left(\sum_{j=1}^k \boldsymbol{\Sigma}_j^{-1} \right)^{-1} \boldsymbol{\Sigma}_i^{-1} \right] \succ 0,$$

and $\boldsymbol{\Phi}$ is positive definite. A symmetric method of moments estimator for \mathbf{T} is

$$\hat{\mathbf{T}} = \frac{\boldsymbol{\Phi}^{-1} \mathbf{A} + \mathbf{A}' \boldsymbol{\Phi}^{-1}}{2}.$$

Bibliography

Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees.

American Journal of Human Genetics 1998; **62**: 1198-1211.

Amos CI. Robust variance-components approach for assessing genetic linkage in

pedigrees. *American Journal of Human Genetics* 1994; **54**: 535-543.

Becker BJ, Wu M. The synthesis of regression slopes in meta-analysis. *Statistical Science*

2007; **22**: 414-429.

Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis

of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:

2537-2550.

Chen H, Manning AK, Dupuis J. A method of moments estimator for random effect

multivariate meta-analysis. *Biometrics* 2012; **68**: 1278-1284.

Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in

family samples. *Genetic Epidemiology* 2013; **37**: 196-204.

Cohn LD, Becker BJ. How meta-analysis increases statistical power. *Psychological*

Methods 2003; **8**: 243-253.

Davies RB. The distribution of a linear combination of chi-square random variables.

Journal of the Royal Statistical Society: Series C (Applied Statistics) 1980; **29**: 323-333.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the

EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*

1977; **39**: 1-38.

DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986;

7: 177-188.

Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting

glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics* 2010;

42:105-116.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing

heritability and strategies for finding the underlying causes of complex disease. *Nature*

Reviews Genetics 2010; **11**: 446-450.

Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a

proper control sample for risk calculations. *Annals of Human Genetics* 1987; **51**: 227-233.

Fisher RA. *Statistical methods for research workers* (4th edition) 1932; Edinburgh: Oliver and Boyd.

Follmann DA, Proschan MA. Valid inference in random effects meta-analysis. *Biometrics* 1999; **55**: 732-737.

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity* 2010; **70**: 42-54.

Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**: 1539-1558.

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. *PLoS One* 2010; **5**: e13584.

Ingels SJ, Pratt DJ, Herget DR, Burns LJ, Dever JA, Ottem R, Rogers JE, Jin Y, Leinwand S. *High School Longitudinal Study of 2009 (HSL:09). Base-Year Data File Documentation* (NCES 2011-328) 2011; U.S. Department of Education, Washington, DC: National Center for Education Statistics. Retrieved Nov. 16, 2011 from <http://nces.ed.gov/pubsearch>.

Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* 2010; **29**: 1282-1297.

Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 1986; **42**: 805-820.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* 2007; **63**: 111-119.

Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 1999; **86**: 929-935.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics* 2008; **82**: 386-397.

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; **13**: 762-775.

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* 2008; **83**:311-321.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13-22.

Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology* 2010; **34**: 60-66.

Lin X. Variance component testing in generalised linear models with random effects. *Biometrika* 1997; **84**: 309-326.

Lindgren CM, Heid IM, Randall JC, et al. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genetics* 2009; **5**: e1000508.

Lipták T. On the combination of independent tests. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* 1958; **3**: 171-197.

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* 2007; **63**: 1079-1088.

Ma Y, Mazumdar M. Multivariate meta-analysis: a robust approach based on the theory of U-statistic. *Statistics in Medicine* 2011; **30**: 2911-2929.

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* 2009; **5**: e1000384.

Manning AK, Hivert MF, Scott RA, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycaemic traits and insulin resistance. *Nature Genetics* 2012; **44**: 659-669.

Manning AK, LaValley M, Liu CT, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. *Genetic Epidemiology* 2011; **35**: 11-18.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 2008; **9**: 356-369.

Montana G. HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics* 2005; **21**: 4309-4311.

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research* 2007; **615**: 28-56.

Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* 2010; **34**: 188-193.

Ott J. Statistical properties of the haplotype relative risk. *Genetic Epidemiology* 1989; **6**: 127-130.

Pankratz VS, de Andrade M, Therneau TM. Random effects Cox proportional hazard model: general variance components methods for time-to-event data. *Genetic Epidemiology* 2005; **28**: 97-109.

Prokopenko I, Langenberg C, Florez JC, et al. Variants in MTNR1B influence fasting glucose levels. *Nature Genetics* 2009; **41**: 77-81.

Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of

genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics* 2009; **2**: 73-80.

Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychological Bulletin* 1988; **103**: 111-120.

Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**: 78-97.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, Lin X. SNP set association analysis for familial data. *Genetic Epidemiology* 2012; **36**: 797-810.

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 1993; **52**: 506-516.

Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM Jr. *The American soldier: adjustment during army life* (Vol. 1) 1949; Princeton: Princeton University Press.

Terwilliger JD, Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity* 1992; **42**: 337-346.

Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; **21**: 1559-1573.

van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**: 589-624.

Wang Y, Chen YH, Yang Q. Joint rare variant association test of the average and individual effects for sequencing studies. *PLoS One* 2012; **7**: e32485.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* 2011; **89**: 82-93.

Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* 2008; **40**: 638-645.

Curriculum Vitae

