2014

Genetic Association Testing of Copy Number Variation

Yinglei Li
University of Kentucky, yinglei2014@gmail.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Yinglei Li, Student

Dr. Patrick Breheny, Major Professor

Dr. Constance Wood, Director of Graduate Studies

Genetic Association Testing of Copy Number Variation

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Yinglei Li
Lexington, Kentucky

Director: Dr. Patrick Breheny and Dr. Arnold Stromberg, Professor of Statistics
Lexington, Kentucky 2014

ABSTRACT OF DISSERTATION

Genetic Association Testing of Copy Number Variation

Copy-number variation (CNV) has been implicated in many complex diseases. It is of great interest to detect and locate such regions through genetic association testings. However, the association testings are complicated by the fact that CNVs usually span multiple markers and thus such markers are correlated to each other. To overcome the difficulty, it is desirable to pool information across the markers. In this thesis, we propose a kernel-based method for aggregation of marker-level tests, in which first we obtain a bunch of p-values through association tests for every marker and then the association test involving CNV is based on the statistic of p-values combinations. In addition, we explore several aspects of its implementation.

Since p-values among markers are correlated, it is complicated to obtain the null distribution of test statistics for kernel-base aggregation of marker-level tests. To solve the problem, we develop two proper methods that are both demonstrated to preserve the family-wise error rate of the test procedure —-a permutation-based approach and a correlation-base approach. Many implementation aspects of kernel-based method are compared through the empirical power studies in a number of simulations constructed from real data involving a pharmacogenomic study of gemcitabine. In addition, more performance comparisons are shown between permutation-based and correlation-based approach. We also apply those two approaches to the real data.

The main contribution of the dissertation is the development of marker-level association testing, a comparable and powerful approach to detect phenotype-associated CNVs. Furthermore, the approach is extended to high dimension setting with high efficiency.

KEYWORDS: Copy number variation, marker-level testing, kernel-base aggregation, permutation, family-wise error rate

Author's signature: Yinglei Li

Date: November 7, 2014

Genetic Association Testing of Copy Number Variation

By
Yinglei Li

Director of Dissertation: Patrick Breheny and Dr. Arnold Stromberg

Director of Graduate Studies: Dr. Constance Wood

Date: November 7, 2014

ACKNOWLEDGMENTS

First I would like to express my deep appreciations and gratitude to my advisors, Dr. Patrick Breheny, for his time, detailed guidance, support and encouragement in every step I made in my research. I greatly benefit from his profound knowledge and scientific insight throughout my PhD study. I sincerely appreciate his instructions and help in assisting me to finish my dissertation. Moreover, Dr. Patrick Breheny helped to develop my research ability, scientific thinking, and analytical skills to a great extent.

I would like to acknowledge my committee members: Dr. Arnold Stromberg, Dr. Richard Charnigo, Dr. David Fardo, Dr. William Griffith for sharing their valuable time and providing me helpful feedback of this dissertation. Your comments always help me thinking more and digging more deeply about my research.

I am very grateful to my friends, Xiang Zhang, Jing Xi, and Shihong Zhu, for their help provided during my dissertation research. Sincere thanks to all of my friends who have made my life full of joy.

My appreciations also go to the Department of Statistics, which provided me with a good environment to study and research.

I would like to thank my parents for your love, encourage, and belief in me; my husband Tongfei, for your love, patience and support during my whole life. My final and most heartfelt acknowledgement must go to my family. I greatly thank my parents, my husband and my two sons for their endless love and supporting my decision

to pursue the Ph.D. and work in research. You are always my constant source of love, happiness, and strength all the years living in a country with a different culture constituted. I made a big step forward and it is my pleasure to take this opportunity to thank all the people contributing to it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Genetic background of Copy number variation

An allele is one of a number of alternative forms of the same gene or same genetic locus that is located at a specific position on a specific chromosome. Genetic variation is the variation in alleles of genes and it occurs both within and among populations. It is brought about by mutation, a change in the chemical structure of a gene. Due to the improved understanding of variation in human DNA and development of high-resolution assays that capable of detecting small segmental genetic alteration, different kinds of genetic variation have been detected. These variations present in an unexpected amount of forms, including single-nucleotide polymorphism, small insertion-deletion polymorphisms, variable numbers of repetitive sequences, and genomic structural variation. Initial genetic variation studies mostly concentrated on single-nucleotide polymorphisms (SNP). A SNP is a DNA sequence variation that occurs when a single nucleotide (A,T,C or G) in a DNA sequence differs between individuals. It is a single base change in the DNA. SNPs are thought to be the most common form of genetic variation. In 2001 it was discovered that there are at least estimated 10 million SNPs commonly occurring within the human population [1]. The international HapMap Project has identified over 3.1 million SNPs across the human genome that commonly exist in the individuals of African, Asian and European ancestry. SNPs have been considered to be the main source of normal phenotypic variation for decades. However, a study by Maher [2] demonstrates that most common single nucleotide polymorphisms associated with complex diseases have small effect size and explained only $2-15\%$ of heritable variation. Increasingly researchers have attempted to discover other type of variation that might account for the remaining fraction of the heritable variation.

Prior to the day of DNA sequencing, very few gene number changes could be detected by a microscope. During the past several years, hundreds of new forms of variation in repetitive regions of DNA have been identified. Two groups of researchers published landmark findings of detection of copy number variation among the genomes of healthy individuals [3, 4]. Subsequent studies using more developed methods expanded the results and provided more evidence for the existence of such variation over large portions of the human genome. The number of copies of a segment of DNA in the genome is referred to as the DNA copy number for that segment. Normally, most individuals have two copies of a given genomic segment of DNA. A different number of copies results when an individual goes through one or more deletions or duplications of a segment. Therefore, by choosing a genome reference sequence, which is considered to have a "normal" number of copies, any individuals with an abnormal number of copies at the same genomic region are defined to possess a CNV at that region.

Copy number variants are now known to be a prevalent form of genetic variation and account for a substantial proportion of genetic variability in human populations. CNVs are categorized in many ways. Some authors subclassify CNVs in terms of the size and frequency. Traditionally, CNV is defined as one type of structural variant in a segment of DNA that is 1 kb or larger [5]. By the improvement of detection technology, shorter segments are able to be detected and now also considered. Some publications defined it with minor allele frequency (MAF). In a recent study [6], $> 90\%$ of CNVs involved copy-number polymorphism that occurs in more than 1% of the given population and $> 80\%$ common CNPs with MAF $> 5\%$. The remaining 10% consists of rare CNVs. This indicates that a large portion of copy number variation come from common polymorphisms. By comparing the human genome ref-

erence sequence with another genome sequence using the fosmid paired-end sequence approach, Tuzun identified 297 sites of potential structure variation between the two sequences, including 139 insertions, 102 deletion, and 56 inversions. [7]. The number of identified CNVs is now increasing dramatically as the detection technologies have improved. To date, approximately 180,000 CNV have been reported in the Database of Genomic Variants (DGV), which records information about identified copy number events. We make no distinction in forms of size or frequency in this dissertation and use the term CNV for all kinds of copy number variation.

More and more studies demonstrate the potentially greater role of CNVs than single base-pair sequence variation. It was estimated that 12% of the genome could be affected by CNV in comparison to $1-2\%$ covered by single nucleotide polymorphisms (SNPs) [8]. Ridon et al. constructed a first-generation CNV map of the human genome after studying 270 individuals. They identified a total of 1,447 copy number variable regions covering 12% of genome in these population. Another research team pointed out that there are approximately 12 CNVs per entire genome on average [3, 4]. But these results were based on the small number of genomes and limited resolution of previous detection methods. Sharp et al. revealed that 61% of the variants identified had not been discovered before [9]. It seems that the number of CNVs is underestimated. By 2006, it appears that 1237 CNVs covering an estimated 143 Mb genomic sequence had been identified, which is a substantial source of genomic variation except SNP [10]. Analysis of the complete DNA sequence from a single individual reported that CNVs account for about 22% of all genetic variation in the individual and 74% of the total DNA sequence variation [11]. It is reasonable to conclude that CNV accounts for an appreciable amount of phenotypic variation.

An increasing number of studies are investigating the impact of copy number variation on various phenotypes. The duplication of the Bar gene in Drosophila was one of the earliest studies on association of CNV with a phenotype. The variation was shown to cause the Bar eye phenotype, which will narrow the eye field of affected flies [12]. Another popular example is the duplication of part or all of chromosome 21 has been discovered to be linked to Down syndrome. More evidence reveals that changes in copy number play an important role in evolution. A study by Stefansson et al. [13] identified a common inversion with frequency of 20% in Europeans that indicates positive selection. They analyzed tens of thousands of samples and investigated that the inversion carrier females have more children than noncarriers. Numerous examples of relevance between CNVs and complex diseases were observed by Redon et al. [14]. CNVs are also found to be associated with Prader-Willi and Angelman syndromes. Sebat et al. discovered association between deletion variation and risk of autism [15]. Walsh et al. reported the contribution to schizophrenia disease and Zhang et al. presented the risk to bipolar disorder [16, 17].

Copy number variants do not necessarily have a negative effect on health. A number of studies provided the evidence that some specific large duplication or deletion do not apparently result in disease [18–21]. In 2004, two groups independently detected the widespread presence of CNVs among the genomes of healthy individuals. [3, 4]. By constructing a targeted bacterial artificial chromosome (BAC) microarray, 119 regions of CNVs were identified, with 73 unreported previously, on a panel of 47 normal individuals representing four different population. The study also demonstrated that segmental duplications occur at hotspots of chromosome rearrangement, acting as causes of normal variants as well as genetic disease [9].

1.2 CNV calling

It has became one of the compelling genetics challenges during the last few years to understand the contribution of copy number variation to the phenotypes such as disease state. As many CNVs are rare, it is difficult to conduct statistically significant association between a single rare CNV and the disease status. Instead, several studies examine the association between phenotype and total CNVs in the whole chromosome. Through case-control testing, these studies identified associations between copy number changes and various diseases, including autism, schizophrenia and bipolar disorder [15–17]. Previous studies of copy-number variation in human populations have largely been restricted to hundreds of individuals and therefore were unable to distinguish variants that are truly rare. More and more recent studies have begun to expand to substantially larger sample size.

Genome-wide association studies (GWAS) in large samples of cases and controls are commonplace and therefore an efficient approach is to carry out association studies involving CNVs using the same data. GWAS is an examination of genetic variation across human genome. It tests markers across the complete sets of genomes of many people to identify genetic association with clinical phenotypes. Researchers use two groups of participants to conduct the study: people with the disease (case) and similar people without the disease (control). The entire genome, which contains millions of genetic variants, is investigated through the case-control testing. If a certain variant is found to be significantly more frequent in people with disease, this type of variation is said to be "associated" with the disease. The associated variant might not directly cause the disease and is considered to be "tagging along" with the actual casual variant. In a typical GWAS, genetic variants are read by SNP arrays, a more popular and common type of DNA microarray data used to detect polymorphisms

within a population. Hence, GWAS typically presents the associations between single-nucleotide polymorphisms (SNPs) and the observed traits. GWAS was first applied in 2005 on patients with age-related macular degeneration(AMD) and two SNPs were detected to influence the risk of disease by case-control testing. Such studies are substantially useful and have already successfully identified hundreds of genetic variations contributing to common, complex diseases. GWA studies often require a large number of samples in order to obtain a reliable signal of risk-SNPs. It was reported by Ehret et al. [22] that the largest sample size was in the range of 200,000 individuals.

There have been several techniques proposed for measuring copy-number variation. Traditionally, large chromosome rearrangements have been detected with G-banded karyotypic analyses or fluorescence in situ hybridization (FISH) using fluorescent probes that bind to part of the chromosomes. With the development of microarray technology, array-based comparative genomic hybridization (aCGH) was one of the widely used techniques. This technology was first introduced as "matrix-CGH" [23]. This technique involves labeling a reference genome and a testing genome with different fluorescence markers, hybridizing to the microarray with genomic clones, and then analyzing the intensity of the hybridization signal for each clone. Then the array-CGH technique has been through a lot of improvements such as BAC Array Comparative Genomic Hybridization, Representational Oligonucleotide Microarray Analysis (ROMA) and Agilent CGH [24] to detect CNVs in human populations.

The CNV studies were restricted to a list of selected candidate genes due to the cost and complexity of CNV assessment. High-throughput single nucleotide polymorphism (SNP)-array technologies provide possibility to investigate copy number

variants (CNVs) in genome-wide scans and specific calling algorithms have been developed to determine CNV location and copy number. High-density single nucleotide polymorphism (SNP) genotyping arrays have recently become more popular for CNV detection and analysis, because the arrays are available for both SNP-based and CNV-based association studies. Furthermore, they provide considerably higher precision and resolution than traditional techniques. Since vast amounts of these data have been already generated during the pursuit of conducting genome-wide association studies (GWAS) among SNPs, it has tremendous advantages and merits in terms of convenience and low expense for studies. Hence it is expected to continue to be the main approach for several years to come. However, there are limitations to using SNP genotyping arrays for CNV detection. One obvious limitation is that SNPs are not uniformly distributed across the genome and are sparse in regions with segmental duplication or deletions [25]. The new-generation SNP arrays, such as the Infinium Illumina Human 1Million probe chip and the Affymetrix 6.0 platform, have now incorporated additional nonpolymorphic (NP) markers to provide more comprehensive coverage of the human genome and thus overcome this limitation. They are proven to be a method for detecting CNVs and now are commonly used to obtain the copy-number measurement recently. In my thesis, I focus here on detection of coy-number variation using raw data from genome-wide single nucleotide polymorphism (SNP) arrays.

In high-density SNP genotyping platforms, a signal intensity measure is summarized for each allele of any SNP. Thus, each marker from SNP arrays consists of two intensity measurements, corresponding to the A and B allele [26, 27]. First, quantile-normalization is required for each probe intensity. After normalization of both intensities, a polar coordinate transformation is applied, generating two parameters R and θ. R refers to the sum of normalized intensities for both alleles and θ

is called as an allelic intensity ratio representing the relative allelic intensity ratio of two alleles. The observed normalized intensity R from a subject is then compared to the expected value of R given neutral copy number, which is computed from linear interpolation of canonical genotype clusters (AA, AB and BB) obtained from a large set of reference samples. Such comparison thus generates a R ratio. Finally, a log transformation is applied to the R ratios. The result, which is called the log R ratio (LRR), is reported along the whole genome for every single marker on the array. LRR serves as a continuous measurement of copy number. A position or a region with no CNV presenting should have LRR equal to zero. Higher LRR intensities are signals of copy number gain at those positions in the test samples and similarly lower LRRs indicate the copy number loss. Analysis of signal intensities across the genome can then be used to identify regions with CNVs [26]. Because of the LRR noise, the drop or increase in LRR might not be obvious. Therefore, it is necessary to apply statistical methods to distinguish the signal from noise.

There are broadly two main estimation goals for CNV detection —- inferring the copy numbers and accurately locating the CNV boundaries. There have been numerous publications proposing methods to segment a genome into various regions of constant copy number, which is called "CNV calling". Some methods involve a penalized likelihood to estimate the breakpoint(CGHseg), using an expectation-maximization-based technique (ChARM) and building hierarchical clustering-style trees along each chromosome (CLAC) [28, 29]. Other papers introduce Bayesian method, the use of a genetic local search algorithm (GA), a wavelet approach(Wavelet), and an extension of the SW-ARRAY algorithm to identify potential breakpoints. Additional proposals include Quantile Smoothing (Quantreg), adaptive weights smoothing (GLAD) and analysis of copy errors (ACE). Hidden Markov model (HMM), circular binary segmentation (CBS), and fused lasso are the most commonly applied techniques. We

present here the description of circular binary segmentation.

First proposed by Olshen et al. [30], circular binary segmentation (CBS) is a non-parametric modified change-point method. As CNVs often extend over multiple markers, it is more resonable to focus on a whole chromosome combining information from neighboring markers than analysis on single one marker at a time. By connecting the two ends of each chromosome, the method offers a nice way to split the chromosome into contiguous regions of constant copy number and model into discrete copy number gains and losses at the same time. Bypassing parametric modeling of the data, it uses a permutation reference distribution to assess the significance of the proposed splits. The selection process, which recursively splits each contiguous segment, is performed until there is no more significant splits among chromosome. Lai et al. reported that CBS performs consistently well compared to other 10 approaches [29].

The main idea behind circular binary segmentation(CBS) is summarized as follows. It uses log R ratios (LRR) as input data and analyzes on each chromosome of a single individual,

1. Joining the first and the last marker of the chromosome, the sequence of LRR intensities become circular.

2. For every potential way of splitting the circle into complimentary arcs, compute the two-sample t-statistics to compare the means between two arcs

3. Segment the circle somewhere that the maximum of test statistics exceeds the critical value

4. Repeat 2-3 steps recursively for testing the change-points until no additional significant segment can be found

The details of the procedure are presented in these two papers [30,31]. The R package DNAcopy is used to conduct the analysis. It is a package of the Bioconductor project and is available at http://www.bioconductor.org/packages/release/bioc/html/DNAcopy.html. It provides the estimates of the mean LRR at every marker among the given genome for output. Those estimates are constant over each arc and therefore give an estimation of the CNV structure. DNAcopy is considered to be one of the best operational algorithm in terms of its sensitivity and FDR criterion for breakpoint detection [28]. Unfortunately, DNAcopy is also one of the slowest algorithms [29].

A number of articles [28, 29, 32–34] have compared numerous methods or algorithms that are adapted for CNV calling. There are no best or optimal algorithm so far. Each algorithm uses different strategy for CNV calling and thus there are obvious variabilities among calling algorithms. The emerging algorithms have different assumptions for CNV detection. In other word, different CNV calling algorithm provides substantially different quantity and quality of CNV calls even for identical raw data. Therefore, the choice of analysis tools is very important for CNV association studies. Some authors have advised using multiple algorithms on the same data to minimize the false discoveries and thus merging call sets from those algorithms could improve sensitivity [33]. Careful normalization of the intensity is required because hybridization signals are very sensitive to experimental noise that might result in even opposite inference. Furthermore, the accurate CNV detection depends on the choice of the array data. Pinto et al. pointed out that some methods are developed specifically for a certain array data and thus perform always better than other algorithms only in such array data [33]. Additionally, many algorithms developed generally per-

form well for identifying large CNVs. Attempting to localize small CNVs even with only one single probe, it results in many false positives. In conclusion, efforts are ongoing to improve methods for CNV calling at the genome-wide level and test for association in large samples of cases and controls.

1.3 Marker-level testing

Two general strategies have been proposed for conducting genetic association studies of copy-number variation. The first approach, which we refer to as variant-level testing, is to do "CNV calling" at the level of each individual first and then carry out association tests of whether individuals with a CNV differ from individuals without a CNV with respect to some phenotype. It is the more popular strategy for testing CNV associations. However, the result of this test mainly depends on the detection of CNV regions in the first step. Moreover, it is still challenging to separate each genome into regions of constant copy number as discussed in last section. Furthermore, there are additional complications for the variant-level approach. One of the big issues is partially overlapping CNVs. Since CNVs are detected person by person, CNVs do not necessarily share the same boundaries. When the sample size gets large, the number of overlapping patterns may be considerable. Whether those represent the same CNV or different CNVs can be a complicated decision. Also, because CNVs do not overlap perfectly among individuals, the association tests are correlated, leading to problems when doing multiple testing. Moreover, it effects the association test when using different threshold to declare a CNV present. If the threshold is too high, true CNVs might be undetected; if the threshold is too low, neutral regions would be defined as CNVs and thus create misclassification and lower power of association testing.

The alternative approach called "Marker-level testing", is also a two-stage procedure, which reverses the order of variant-level testing. Since variant-level testing presents a number of difficulties, marker-level testing becomes an attractive alternative. In the first stage of marker-level testing, we carry out association testing at the level of the single marker between raw intensity measurements and phenotype of the null hypotheses "H_{0i}: the ith marker is not associated with the ith phenotype". Here intensity is a continuous measurement of copy number at every genetic marker. Furthermore, we could get the association test results using a linear regression model if the phenotype is also a continuous variable. Note that we do not involve specific CNV calling in this step and thus we do not have to deal with difficulties from "CNV calling". The key point here is that, because the data is noisy, it is virtually impossible to identify CNV associations from a single marker. Since copy number variants span over multiple markers in a sufficiently high-density array, the presence of a single CNV that affects the phenotype will elevate the test statistics for several nearby markers. This is the motivation for the second stage of marker-level testing. To carry out inferences regarding CNVs by pooling test results across neighboring markers to determine CNV regions associated with the phenotype is the main job of stage II. By conducting a multi-locus association test along the chromosome, combining neighboring p-values, we can detect a region for copy number-phenotype association with low p-values in close proximity to each other marker. It requires a systematic method for pooling information across the neighboring hypothesis tests to identify the regions in which low p-values are aggregated. In the dissertation, I develop a kernel based approach on p-values to identify CNVs associated with a phenotype by aggregating p-values and show it has a well-controlled false positive rate and high power.

The negative log transformation of the p-values are shown in Figure 1.1 for every marker along a part of chromosome. The p-values are the results from the first step

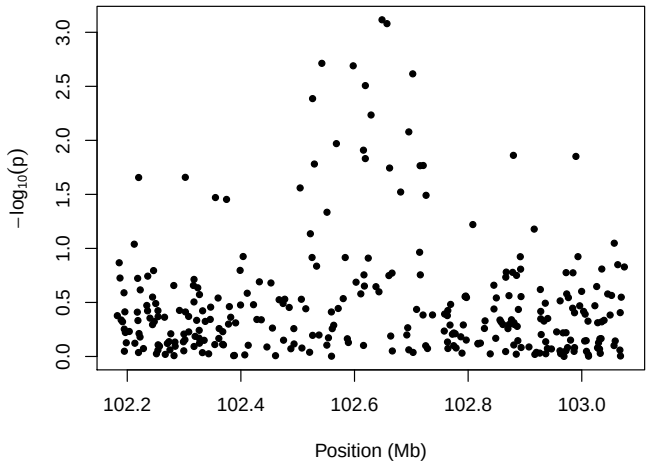

Figure 1.1: Illustration of marker-level testing. The $-\log_{10}(p)$ values on each position along the partial chromosome based on the marker-level tests.

of marker-level testing, which test associations between copy number intensity and phenotype at every marker. From the plot, we spot the cluster of small p-values between 102.5 and 102.7 Mb. This region with so many low p-values in close proximity to one another suggests an association between the phenotype and the copy number variation. In addition, marker-level testing, the main idea of which is illustrated in the figure, avoids the complications of overlapping problems in variant-level testing.

Both of the above statistical technologies consist of two stages. This might result in power loss in the second stage since those approaches risk losing information in the the first stage. The type of information lost by each approach is different and hence it is strongly implicated for the power comparing between those two approaches. A recent paper by Breheny et al., in which CBS was used for segmentation in both approaches, demonstrates that variant-level testing has greater power to detect association involving large, rare CNVs while marker-level testing has more power to detect small, common CNVs [35].

1.4 Outline of the dissertation

The remainder of this thesis is organized as follows. In Chapter 2, a kernel-based aggregation of marker-level association test is developed in detail. Since p-values between markers are correlated with each other, the exact distribution becomes complicated. In the following 2 chapters, two different computational approaches are proposed to overcome the difficulty. I introduce a permutation procedure in chapter 3 and correlation method in chapter 4. In these two chapters, I also illustrate the preservation of type one error and apply both approaches to both simulated and real data. Finally, I summarize the results and discuss future directions in Chapter 5.

Chapter 2 Kernel-based Aggregation Method for Marker-level Association Test

2.1 Introduction

Recent advances in genome, such as the completion of human genome sequence and rapid improvements in SNP genotyping technology, have accelerated the process of locating candidate genes. The development of statistical and computational strategies on hundreds of loci have investigated relationships between genome variation and phenotypic variation. In spite of such advances, no comprehensive, well-powered approach has been published to identify genes and loci that contribute to common disease. It has been receiving considerable attention to detect and locate CNV markers over the past several years. Association test is broadly considered and highly accurate to identify disease susceptibility genes related to complex disorder [36]. The choice of association test is one of the main and important factors for a successful association study. Before the emergence of large-scale association studies, many study of human genetics has suffered from the problem of inadequate statistical power, which results in low rates of successful replication among reported significant associations. It is always worthwhile to choose a good statistical method to maximize the statistical power and preserve well-controlled false positive rate.

Genome-wide association studies (GWASs) were made feasible in the late 2000s. Because of the enormous size of the data sets, GWASs have tended to use simple statistical procedures for one gene at a time throughout the genome. This single-locus testing detects one gene at a time and is more suitable to study a susceptibility gene with strong main effect on complex disorders. Evaluating one gene at a time

focuses on its marginal effect on disease. Of course, there is no problem if focusing on only one particular hypothesis. The marginal or individual p-values does not incorporate any information about the dependence structure between these p-values. As many genes are being tested simultaneously, keeping the significance threshold at the conventional value of 0.05 would lead to a large number of false positive significant results. Such classic nominal significance-threshold framework is inappropriate for such studies. For example, if $\alpha = 0.05$ is the nominal significance rate for one marker and $n = 100$ independent markers are tested at the same time, then false-positive results will be obtained at a frequency of $1 - (1 - \alpha)^n = 0.99$. It means that there is a chance of greater than 99% that one or more markers are shown to be significantly related to disease. This is obviously an unacceptable rate. Such problem involving multiple testing and its effect on the genomewide type I error is the subject of a ongoing debate [37]. There have been proponents for an alternative approach to multiple testing adjustments. The traditional and the most widely used method is standard Bonferroni correction. The Bonferroni correction sets the critical significance threshold at α divided by the number of tests. For example, with 20 tests and $\alpha = 0.05$, you'd only reject a null hypothesis if the p-value is less than 0.0025. The Bonferroni correction tends to be a bit too conservative. We calculate the probability of observing at least one significant result to be 0.0488 when assuming that all tests are independent of each other. In practical applications, that is often not the case. Depending on the correlation structure of the tests, the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives.

Identifying regions of genome or candidate genes that are correlated to contribute to disease is difficulty because complex traits presumably arise from multiple interacting genes with rather small effect. As there are many susceptibility loci, each single gene will have only a small effect and cannot easily be detected by single-locus method. Therefore, it would be appropriate to consider and analyze sets of marker loci jointly

for genomewide association analysis rather than marker-by-marker approach that completely ignores the possible interactions between susceptibility genes. In recent years, multi-locus method is widely used for association study to localize the disease-related genes [38] since a high density of markers and high-resolution microsatellite maps are available. Investigating association between marker genotypes and disease phenotypes for multiple markers will capture more information regarding the total combined effects of all disease genes and therefore increase the statistical power for disease gene detection compared to single-locus inference that analysing one locus at a time. Researchers have developed different kinds of multi-locus association analysis. For instance, there are multi-factor dimensonality reduction (MDR) by Ritchie et al. [39], statistic combination tests by Hoh et al. [38] and p-value combinations by Zaykin et al. [40–42].

Our research focuses on the p-value combination methods. p-value combination began with Fisher's product p-value method or the sum of log scale of p-values (PPM) [43]. Later, other p-value combination methods such as the sum p-value method(SPM) [44] and the minimum p-value method (MPM) [45] were developed. Then p-value truncation was introduced by Wilkinson [46] and extended into truncated product p-value method [40] and recently the rank truncated product p-value method [41, 47]. There is no uniformly most powerful method of combining p-values [40]. In this thesis, I am aimed to develop a new and modified p-value combination method for powerful multi-locus association scans and apply it to large-scale simulation studies and real data analysis.

2.2 Kernel-based aggregation of marker-level association tests

I introduce a two-stage procedure for an association study to locate disease susceptibility CNVs related to complex traits. Let n denote the sample size and J denote the total number of markers. Consider a study region that contains J markers at different positions. Here I use i to index subjects and j to index markers with positions ℓ_j used to denote the location of marker j along the chromosome. Suppose our analysis data consist of a set of marker genotypes together with phenotypic trait values measured on each individual $(\boldsymbol{X}^i, y_i)$, $i = 1, 2, ...n$, where $\boldsymbol{X}^i = (X_{i1}, ..., X_{iJ})^T$ are the copy number intensity of every marker for subject i and y_i is the phenotype for subject i. Here X_{ij} represents the intensity measurement for subject i at marker j. In the first stage, we conduct J association tests for every single locus under the null hypothesises H_{0j}: the jth marker is not associated to the phenotype, $j = 1, ..., J$. I refer to the location at which a test statistic is computed as an analysis point. Then a series of p-values are calculated from the association test between intensity and phenotype for every marker. The association from the single-locus tests is the marginal effect of each locus in the genome ignoring inter-marker association. The key to detect CNV is the detection of significant association between phenotypic trait values and the markers or intervals in a genetic map, which leads to the second stage.

In the second stage, a multi-locus association test combining multiple neighboring p-values is performed. Compared to single-locus tests, multi-locus tests may increase testing power by including the combined effect of disease-associated loci and inter-marker relationship. Consider a window with multiple markers. Define the the central marker as anchor locus, which is denoted as ℓ_0. The anchor marker is used to capture disease genes. Let h denote bandwidth and construct a window by simultaneously considering h before and after the anchor. I apply a kernel-weighted moving average

for each window across our study region. Sliding windows formed scanning the entire study region. Scanning all loci from the starting marker to the end marker, the whole study region will be divided into overlapping windows as we shift the anchor locus. Within each window, consider the local average for anchor ℓ_0, the center of a window.

$$T(\ell_0) = \frac{\sum_j t_j K_h(\ell_j, \ell_0)}{\sum_j K_h(\ell_j, \ell_0)}, \tag{2.1}$$

where $t_j = f(p_j)$ is a function of the p-value for marker j. The smoothing parameter h, which defines the bandwidth of the kernel, controls the size of the neighborhood around the center location ℓ_0. Every marker within the window may contribute to the identification of disease genes, and the extent of each marker contribution is considered by assigning proper weights to markers within a window. The kernel function K controls the weight given to every p_j in the neighborhood based on how far away marker j is along the chromosome from the target location ℓ_0. The higher weights are assigned to the markers closer to the anchor and lower effects for remote marker loci. The smoothing parameter h meanwhile effects the bias-variance tradeoff. A larger bandwidth will decrease variance by pooling p-values among a boarder region but introduce more bias because considering more extra test results beyond the boundary of a CNV.

One could apply (2.1) at any arbitrary location ℓ_0. We restrict our consideration to a finite set of aggregations $\{T_j\}$ from reasonable locations at which a marker is present and the bandwidth does not exceed the borders of chromosome. Furthermore, the transformation of p-values is chosen such that low p-values produce large values of t_j, leading to significance testing based on the statistic $T = \max_j\{T_j\}$.

In this section, we describe in detail the choice of kernel K_h and transformation $f(p_j)$, as well as the issue of the direction of association for signed tests.

2.2.1 Choice of kernel

In nonparametric analysis, the goal is to smooth the data in some way and estimate a curve. Kernel-based approaches are an important class of smoothing methods. The main idea behind the method is to assign location weight based on each observation and then aggregate them to yield an overall local average. Additional details of this method and its application to CNV association studies will be discussed here.

There are usually two primary choices with regard to the kernel: kernel function and bandwidth selection. First, let's consider the shape of the kernel. Consider two frequently used kernel density functions, flat ("boxcar") kernel and the Epanechnikov kernel, are considered in our study:

$$\text{Flat(boxcar)}: \quad K_h(\ell_j, \ell_0) = \begin{cases} 1 & \text{if} \quad |\ell_j - \ell_0| \leq h \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$$\text{Epanechnikov}: \quad K_h(\ell_j, \ell_0) = \begin{cases} \frac{3}{4}\left\{1 - \left(\frac{\ell_j - \ell_0}{h}\right)^2\right\} & \text{if} \quad |\ell_j - \ell_0| \leq h \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

The Epanechnikov kernel would seem to be more attractive, as it gives higher weight to markers close to the anchor location, and negligible weight to remote markers where bias is a larger concern.

Besides varying the shape of kernel, the quality of the kernel estimate depends more on the definition of its bandwidth. Since the aggregation function (2.1) considers the weighted average of a few markers within the bandwidth of the center marker, it would be interesting to consider two definitions of bandwidth, which we

refer to as *constant width* and *constant marker* (these concepts are named "metric" and "adaptive" bandwidths, respectively, in kernel smoothing literature). In the constant width approach, the width h of the kernel is constant in functions (2.2)-(2.3). Meanwhile the number of markers for every kernel window varies the target location ℓ_0 changes, thereby suffering from fluctuating variance. In contrast, the constant marker approach provides constant number of markers for each kernel window. But meanwhile it produces various range of the kernel, thereby suffering from bias. Specifically, $h_k(\ell_0) = |\ell_0 - \ell_{[k]}|$, where $\ell_{[k]}$ is the location of the kth closest marker to target location ℓ_0. Simulation results regarding the benefits and drawbacks of these various kernels are shown in the next chapter.

As a matter of reference, the flat, constant marker kernel is similar to the simple moving average, although not exactly the same. The boxcar kernel for constant marker method assigns the same weight of 1 to the effective markers which are within the h nearest neighbors to ℓ_0 and assigns 0 otherwise. For example, consider the following illustration.

Suppose $h = 3$. We keep varying the target location, ℓ_0 to get the aggregation for each marker. Unlike the moving average, constant marker kernel looks for the nearest neighborhood to the target location. For example, at ℓ_3, the three nearest neighbors are $\{p_1, p_2, p_3\}$, while at ℓ_4, the three nearest neighbors are $\{p_4, p_5, p_6\}$. Thus, combinations such as $\{p_3, p_4, p_5\}$ are not considered by the kernel approach. One obvious benefit of this method is to exclude the aggregation over inappropriately disperse regions of the chromosome.

2.2.2 Transformation of p-values

As suggested in (2.1), directly pooling p-values is not necessarily optimal. We consider various transformations of p-values in a way that low p-values lead to high values of $t_j = f(p_j)$ and hence the statistic $T = \max_j\{T_j\}$ of association testing. Such transformations might better discriminate true association from noise. Three main transformations, uniform, Gaussian, and logarithmic, are considered as follows:

$$\text{p}: \quad t_j = 1 - p_j \tag{2.4}$$

$$\text{Z}: \quad t_j = \Phi^{-1}(1 - p_j) \tag{2.5}$$

$$\log: \quad t_j = -\log p_j, \tag{2.6}$$

where the text to the left of the equation is the label with which we will refer to these transformations in later figures and tables.

As mentioned in Section 2.1, all these three transformations have a long history in the field of combining p-values methods. Ronald Fisher first proposed the Fisher combination test, which is based on the the average of $\log p$ values or equivalently, the log of the product of the p-values [48]. An alternative procedure developed by Samuel Stouffer [49] depends on the sums of normal-transformed p-values $Z_i = \Phi^{-1}(1 - p_i)$. Finally, (2.4) was studied and derived by Edgington [44]. Several other researchers have followed upon and developed these methods [50–53]. Throughout this thesis, the majority of work will focus on these thee scales— uniform, Gaussian, and logarithmic. Each of these methods has its respective advantages and has proven practical and valuable in different fields. There is no uniformly most powerful method of combining p-values [40]. Comparisons of the uniform, Gaussian, and logarithmic transformations by simulation are shown in the later chapter.

In the p-value combination literatures, tests were assumed to be independent for the convenience of theoretical development. However, this assumption is too stringent for many practical application. If p-values within a window are statistically dependent, a problem common to all of these methods is the difficulty of determining or approximating the distribution of test statistic under an appropriate null hypothesis. More recently, different computational algorithms have been proposed to generate null distribution with dependent p-values, such as permutation, bootstrap, and Monte Carlo [54].

Moreover, since marker-level testing does association tests for every marker among the chromosome first and then locates the CNV, the borders of the CNVs are unknown, as is the appropriate set of p-values to combine. Consequently, we must calculate many different combinations $\{T_j\}$, which yield partially overlapping sets and contain p-values of the same markers in different combinations. Clearly, the resulting test statistics $\{T_j\}$ will not be independent. This is a concern that must be addressed for further studies. The implications of such concerns are addressed in chapter 3 and 4.

2.2.3 Direction of association

In (2.1), the association test between intensity and phenotype at each marker, is unrestricted and could arise from any kind of association test. Some association tests such as z tests and t tests have a direction associated with them, while others such as χ^2 tests and F tests do not. If the direction is available along with the test, it is advantageous and valuable to incorporate this direction into the analysis of multi-locus association tests as we will see in Section 3.4.

Let s_j denote the direction of association at marker j. For example, in a case control study, if intensities are higher for cases than controls at marker j, then $s_j = 1$. Otherwise, at markers where CNV intensities are higher for controls than cases, $s_j = -1$. The signs are arbitrary; their purpose is to reflect the fact that switching directions of association are inconsistent with the biological mechanism being studied — an underlying, latent CNV that affects both phenotype and intensity measures — and thus likely to be noise. Considering this direction should diminish noise and thus improve CNV detection. I introduce here extensions of the transformations presented in Section 2.2.2 that include the direction of association.

When s_j is available, I adjust the three transformations from 2.2.2 as follows:

$$\text{p}: \quad t_j = s_j(1 - p_j) \tag{2.7}$$

$$\text{Z}: \quad t_j = \Phi^{-1}\left(\frac{1 + s_j(1 - p_j)}{2}\right) \tag{2.8}$$

$$\log: \quad t_j = -s_j \log p_j. \tag{2.9}$$

All of these transformations have the same effect as mentioned in Section 2.2.2: when $p_j \approx 0$ and $s_j = 1$, $t_j \gg 0$; when $p_j \approx 0$ and $s_j = -1$, $t_j \ll 0$; and when $p_j \approx 1$, $t_j \approx 0$ regardless of the value of s_j. In other words, all the test results combine to give an aggregate value $T(\ell_0)$ that is large in absolute value only if the test results have low p-values and are consistently in the same direction.

2.2.4 Summary

The kernel-based aggregation method described above provides a nice way to test for CNVs associated with the phenotype. First, we obtain a list of p-values for every marker through multiple hypothesis tests. Then we combine and transform p-values to yield a finite set of aggregations $\{T_j\}$. The final significance testing is based on the

statistic $T = \max_j\{T_j\}$. Here as described above, different choices of kernel, transformations of p-values and the direction of association test will effect the statistic and thus effect the test power. The power comparisons through different choice will be presented in the later chapter through simulation.

To determine the significance of $\max_j\{T_j\}$, we must estimate its null distribution. Since the p-values are not independent among markers, the estimate of the exact null distribution gets complicated. In chapter 3 and 4, we propose two approaches — one based on a permutation approach and the other based on estimating the correlation structure directly — to give an estimate of the null distribution. Moreover, we apply these approaches to the simulated and real data.

Chapter 3 Family Wise Error Rate Control for Permutation Method

3.1 Introduction

In genomics, especially genetic association studies, it is common for tens of thousands of genes or genetic markers to be measured. For our marker-level association testing, the first step involves a large number of markers being testing simultaneously. As discussed in Section 2.1, multiple hypothesis testing is a common problem in such genomewide association studies. If one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected by chance alone may be unexpectedly large. In our studies, it is of great interest to determine the phenotype-associated CNV region that spans several markers in the second step. And we consider the test statistic containing a list of p-values in a window for significant result. Thus such problem involving multiple testing arises and should be considered and paid more attention since the decision is based on the dependent structure of corresponding marginal p-values.

Failure to consider the effects of multiple comparisons would result in an abundance of false positive results. To deal with the multiple testing problem, we must extend the idea of type I error to acknowledge multiple testing. The traditional or classical criterion for error control is the familywise error rate(FWER), which is defined as the probability of rejecting one or more true null hypothesis. A procedure is said to control FWER in the strong sense if FWER control at level α ($FWER \leq \alpha$) regardless of which subset of hypotheses is true. Similarly, if the FWER control at level α only when all null hypothesis are true, it is defined to control FWER in the weak sense. We are then $1 - \alpha$ confident that there are no false discoveries among

the rejected hypothesis. Another alternative criterion for error control is the false discovery rate(FDR), which is the expected proportion of falsely rejected hypotheses. It was first introduced by Benjamini and Hochberg when proposing a step-down procedure to control FDR for independent structures [55]. It was later shown that this procedure also controls FDR for certain dependence structures [56]. FDR is equal to FWER when all null hypotheses are true. When the vast majority of the null hypotheses are true, as the case in association studies, it is common to focus on FWER for the sake of simplicity. However, it would more appealing to use FDR in situations where there are a large number of false null hypotheses involved [57].

With a high density of markers, our test statistics as described are highly correlated for multiple testing. The assumption of independence among tests is strongly violated. Many different computational approaches have been developed to overcome this difficulty. One alternative approach is permutation testing approach [45, 58]. Permutation tests shuffle the phenotype values among subjects a number of times and keep the order of genotype set to create permuted data sets that have random genotype-phenotype associations. Monte Carlo procedure was proposed by Zaykin [40] and a direct simulation approach was advised by Seaman and Müller-Myhsok [59]. Permutation resampling procedure was considered in this chapter since it is traditional and popular. The empirical joint distribution of the test statistics using such permuted data serves as the reference null distribution to determine the CNV-phenotype association.

In this chapter, I will describe the permutation method, which is applied to our kernel-based aggregation of marker-level test. Then I will illustrate that this approach controls FWER through simulation and proof. Applying such method through sim-

ulation, I compare many aspects of the kernel-based aggregation association test statistic.

3.2 Significance testing and FWER control

3.2.1 Exchangeability

In any analysis that involves aggregating marker-level test results, we must be able to detect and quantify the significance of regions like those depicted in Figure 1.1. This is not trivial, however. As we described in Section 2.2.2, The fact that p-values among markers within a window are statistically dependent greatly increases the difficulty of estimating the exact null distribution. Thus this dependence introduces a lack of exchangeability between test results, which complicates matters and causes various naïve approaches to fail. In this section, I compare three approaches that I tried during my research and illustrate the consequences of non-exchangeability in testing the significance of a region with a preponderance of low p-values.

One approach, suggested in [35], is to use circular binary segmentation (CBS; implemented in the `R` package `DNAcopy`). This method aggregates neighboring p-values by calculating the two sample t-test statistic comparing the mean intensity of a given region with that of the surrounding region. The significance of this test statistic is quantified by comparing it to the distribution of maximum test statistics obtained by permuting the $\{p_j\}$ values [30, 31]. However, the main assumption of this approach is that the test results $\{p_j\}$ are exchangeable which is the justification for permuting them.

Alternatively, we may use the kernel-based method described in Section 2.2.1 to aggregate the neighboring test results, thereby obtaining $T_{\max} = \max_j\{T_j\}$. One possible approach to generate the null distribution of $T_{\max}$ is to rely on Monte Carlo integration based on the fact that, under the null hypothesis of no association, all p-values follow a uniform distribution. Thus, for any choice of transformation and kernel in (2.1), we may generate an arbitrary number of $\{T_j\}$ under the null and then yield independent draws $\{T_{\max}^{(b)}\}_{b=1}^{B}$ from the null distribution function F_0 of $T_{\max}$. Then the estimate $\hat{F}_0$ is obtained using the empirical CDF of those draws $\{T_{\max}^{(b)}\}_{b=1}^{B}$. Through this approach, we apply a test for the significant presence of a CNV-phenotype association through the calculation of $p = 1 - \hat{F}_0(T_{\max})$. The crucial assumption here is that, under the null, the p-values among markers are independent and so are $\{T_j\}$.

An alternative to generate the null distribution and quantify the significance of $T_{\max}$ is the permutation approach that is proposed and described fully in Section 3.2.2. By permuting the phenotype prior to aggregation of the marker-level tests, it creates the independence between intensity and phenotype among markers for each permuting. Thus, using the empirical CDF of an arbitrary number draws $\{T_{\max}^{(b)}\}_{b=1}^{B}$, we would obtain the estimate $\hat{F}_0(T_{\max})$.

Consider a genomic region in which individuals may have a CNV. The purpose of the analysis is to detect and locate such a CNV if it is associated with a particular phenotype. Thus, the null hypothesis for our association test may hold in one of two ways: (1, "No CNV") no individuals with CNVs in that region are present in the sample, or (2, "No association") individuals with CNVs are present in the sample, but the CNV does not affect the disease and thus dose not change the probability of develop-

ing the phenotype. The preservations of type I error of the three methods discussed above under each type of null hypothesis is shown in Table 3.1. It demonstrates that while all three methods have the proper type I error rate in the 'No CNV" setting, only the permutation approach preserves the correct type I error in the case where a CNV is present, but not associated with the disease ("No association"). It's easy to see that p-values are independent for all methods under null hypothesis 1 ("No CNV"). When a CNV is present but not related to the disease for null hypothesis 2 ("No association"), it is still true that the marginal distribution of each p_j is Uniform(0,1) for each marker. This phenomenon is also illustrated graphically in Figure 3.1.

Table 3.1: Preservation of Type I error for three methods with nominal $\alpha = .05$ in two possible settings for which the null hypothesis holds. The simulated genomic region contained 200 markers, 30 of which were spanned by a CNV. The CNV was present in either 0% or 50% of the samples, depending on the null hypothesis setting. A detailed description of the simulation data is given in Section 3.4.

	Circular binary segmentation	Kernel Monte Carlo	Kernel Permutation
No CNV	0.05	0.06	0.06
No Association	0.20	0.54	0.06

Table 3.1 and Figure 3.1 demonstrate that CBS and kernel Monte Carlo are not guaranteed to preserve the type I error in all settings. Exchangeability is very crucial to be considered when estimating the null distribution. I also make the following additional observations from comparison results: (1) The CBS approach is somewhat more robust to the exchangeability issue than the Monte Carlo approach; *i.e.*, its type I error rate is not as badly violated. (2) The data simulated here for the "no association" setting are a little bit exaggerated: the CNV was present in 50% of the population and the signal to noise ratio was about twice as high as that typically

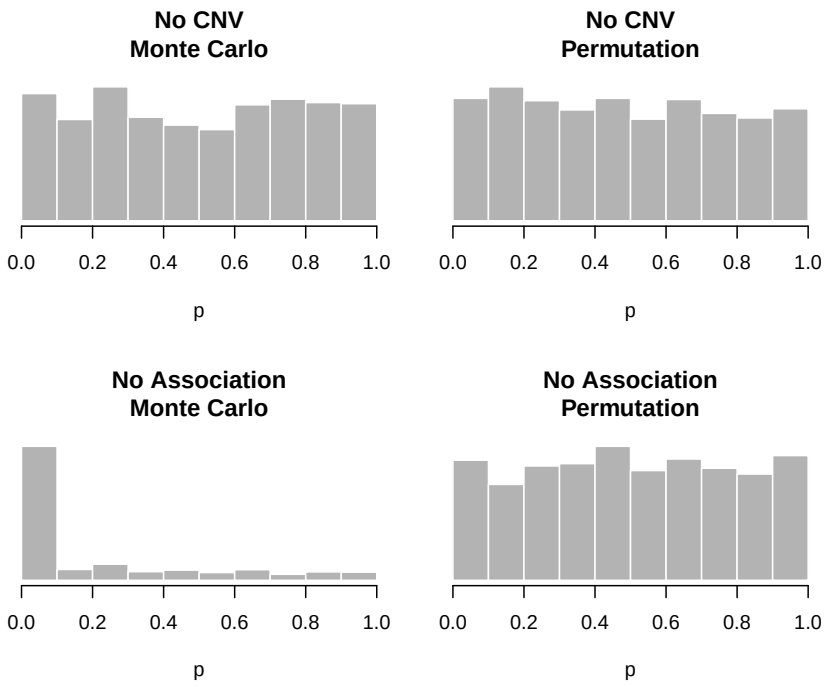

Figure 3.1: Ability of Monte Carlo and Permutation approaches to maintain family-wise error rate under the two null scenarios. The implementation of CBS provided by `DNAcopy` does not return p-values (only whether they fall above or below a cutoff), and thus could not be included in this plot.

observed in real data. In more realistic settings, the violation of type I error rate will be not nearly as severe. (3) Circular binary segmentation was developed for the purpose of detecting CNVs, not aggregating marker-level tests, and thus its failure to preserve the family-wise error rate in this setting is in no way a criticism of CBS in general.

3.2.2 Permutation approach

The concept of permutation tests was first proposed by Fisher(1935). Good(1994) provided an introduction to the theory of permutation testing. A summary of the theory of permutation tests can be found in Lehmann (1986). Permutation tests are a

class of widely-applicable non-parametric tests. They use random shuffles of the data to estimate the correct distribution of a test statistic under a null hypothesis. They provide valid tests with the advertised Type I error, although they are more computationally intensive than standard statistical tests. Permutation tests are widely used in genetics and genomics. They are especially useful when we have insufficient information about the distribution of the data, are uncomfortable making assumptions about the distribution, or if the distribution of the test statistic is not easily computed. They are used in candidate-gene and genome-wide association studies, as well as in family-based association tests.

In our case of CNV-association testing, a permutation test gives a simple way to compute the sampling distribution for the test statistic, under the null hypothesis that a set of genetic variants has absolutely no effect on the outcome. Permutation involves randomly repeated "shuffling" of the phenotype trait values and thus creates many samples under the null hypothesis to estimate the sampling distribution of the test statistics. I formally define the kernel permutation method introduced in Section 3.2.1 and show that it preserves family-wise error rate for the problem of CNV association testing. In this section below, I will fully describe this approach and prove that it preserves type I error and controls FWER theoretically.

For a given set of test results $\{p_j\}$ and $t_j = f(p_j)$, consider the statistic

$$T_{\max} = \max_j \{T_j\}, \tag{3.1}$$

where T_j is the kernel-based aggregation of t_j in a window. If signs of the tests are available , with results $\{p_j, s_j\}$, we use $T_{\max} = \max_j\{|T_j|\}$.

To estimate the null distribution of $T_{\max}$, we use a permutation approach, generating up to $n!$ unique draws $\{T_{\max}^{(b)}\}_{b=1}^{B}$ from the permutation distribution of $T_{\max}$. The procedure is as follows. At any given iteration, draw a random vector of phenotypes $y^{(b)}$ by permuting the original vector of phenotypes. That is, reassigning each phenotypic trait to a new individual while retaining the individual's genetic intensities. Under the null of no association between intensities and phenotypes, randomly shuffling the phenotypic values across individuals will not alter the distribution of the test statistic. Next, carry out marker-level tests of association between the original CNV intensities and the permuted vector of phenotypes, obtaining a vector of permutation test results $\{p_j^{(b)}\}$. Finally, apply the kernel aggregation procedure described in Section 2.2.1 to obtain $\{T_j^{(b)}\}$ and $T_{\max}^{(b)}$. The entire procedure is repeated B times. In this way, we reach the independence between intensities and phenotypes for each iteration such that all permutated results $\{T_{\max}^{(b)}\}_{b=1}^{B}$ are equally likely and exchangeable. Computing appropriate test statistics from each shuffling, we are essentially sampling from a null distribution corresponding to no association between intensities and trait values. We may then use the empirical CDF of these draws from the permutation distribution of $T_{\max}$ to obtain the estimate $\hat{F}_0$. Thus, we obtain a global test for the significant presence of a CNV-phenotype association based on $p = 1 - \hat{F}_0(T_{\max})$. By preserving the correlation structure of the original CNV intensities, this approach does not rely on any assumptions of exchangeability or independence across neighboring markers, and is thereby able to preserve the type I error rate of the testing procedure, unlike the other approaches described in Section 3.2.1. I now formally present this result, the proof of which appears as below.

Theorem 1. *Let H_0 denote the hypothesis that the phenotype, y_i, and the vector of CNV intensities, x_i, are independent. Then, using the permutation approach described*

above for any of the kernel aggregation approaches in Section 2.2.1, for any $\alpha \in (0, 1)$,

$$\Pr(\textit{Type I error}) \leq \alpha. \tag{3.2}$$

Proof of Theorem 1. Let $\mathcal{P}$ denote the set of all possible permutations of $\{y_i\}$, F_0 the CDF of $T_{\max}$ over $\mathcal{P}$, and F_0^{-1} its generalized inverse. Also, let $\phi(\mathbf{X}, \mathbf{y}) = 1$ if $T_{\max}(\mathbf{X}, \mathbf{y}) > F_0^{-1}(1 - \alpha)$ and 0 otherwise.

Now, note that under the null hypothesis that $\mathbf{x}_i$ and y_i are independent,

$$\begin{aligned} P(\mathbf{X}, \mathbf{y}) &= \prod_i P(\mathbf{x}_i, y_i) \\ &= \prod_i P(\mathbf{x}_i) P(y_i) \\ &= P(\mathbf{X}, \mathbf{y}^*) \end{aligned}$$

for all $y^* \in \mathcal{P}$. Thus, $\mathbb{E}_0 \phi(\mathbf{X}, \mathbf{y}^*)$ is a constant for all $\mathbf{y}^*$ and

$$\begin{aligned} \mathbb{E}_0 \left\{ \phi(\mathbf{X}, \mathbf{y}) \right\} &= \frac{1}{n!} \sum_{\mathbf{y}^* \in \mathcal{P}} \mathbb{E}_0 \phi(\mathbf{X}, \mathbf{y}^*) \\ &= \mathbb{E}_0 \frac{1}{n!} \sum_{\mathbf{y}^* \in \mathcal{P}} \phi(\mathbf{X}, \mathbf{y}^*) \\ &\leq \alpha, \end{aligned}$$

where the term inside the expectation in the second line is less than or equal to α for all $\mathbf{X}$ and $\mathbf{y}$ by the construction of the test. □

The permutation test is guaranteed to have the correct desired false positive rate (Type I error) regardless of the distributional characteristics of the data at hand. It is worth pointing out that the above theorem is proven for the case in which all permutations of $\{y_i\}$ are considered. In practice, as it is usually impractical to consider all permutations, one can generate only a random samples of these permutations from the set of all permutations of the data. However, by the law of large numbers, the above conclusion still holds approximately, and may be made as precise as necessary

by increasing the value of B, the number of permutations evaluated. For the numerical results in Section 3.4, we use $B = 1,000$.

The global test above aims to quantify and represent compelling evidence for a CNV-phenotype association among the whole chromosome. However, it is of limited practical benefit in the sense that it does not indicate the location of the associated CNV. Thus, we could consider the following equivalent marker-level test: declare significant evidence for the presence of a CNV-phenotype association at any marker for which $T_j > F_0^{-1}(1-\alpha)$. Below, we state the corollary to Theorem 1 for the kernel permutation method, viewed as a multiple testing procedure for each marker.

Corollary 1. *Let H_{0j} denote the hypothesis that the phenotype, y_i, and the CNV intensity at marker j, X_{ij}, are independent. Then, under the global null hypothesis that y_i is jointly independent of $\{X_{ij}\}$, for any $\alpha \in (0,1)$,*

$$\Pr(\textit{At least one Type I error}) \leq \alpha \tag{3.3}$$

using the permutation approach described above and $T_j > F_0^{-1}(1-\alpha)$ as the test function for H_{0j}. In other words, the testing procedure described above controls the FWER in the weak sense at level α.

It is worth noting that the procedure above controls the FWER only in the weak sense — in other words, that it limits the probability of a false declaration of a CNV only under the global null hypothesis that there are no CNVs associated with the outcome. Typically in multiple testing scenarios, this is undesirable and strong control is necessary. However, in the case of CNV-phenotype association, strong control is impractical, as it would imply that a method not only identifies CNV-phenotype associations, but can perfectly detect the genomic boundary of any associated CNV.

This is an unrealistic requirement; in practice, there is no way to prevent the possibility that a detected CNV-phenotype association may spill over beyond the boundary of the CNV.

3.3 Gemcitabine study

In this section we describe a pharmacogenomic study of gemcitabine, a commonly used chemotherapeutic agent in many kinds of cancer. I begin by describing the design of the study [35], then analyze data from the study using the proposed kernel-based aggregation method by permutation. This data will also be used to create spike-in simulated data sets for power comparison of simulation studies in Section 3.4.

The gemcitabine study was carried out on the Human Variation Panel, a model system consisting of cell lines derived from Caucasian, African-American and Han Chinese-American subjects (Coriell Institute, Camden, NJ). Gemcitabine cytotoxicity assays were performed at eight drug dosages (1000, 100, 10, 1, 0.1, 0.01, 0.001, and 0.0001 uM) [60]. Estimation of the phenotype IC50 (the effective dose that kills 50% of the cells) was then completed using a four parameter logistic model [61]. Marker intensity data for the cell lines was collected using the Illumina HumanHap 550K and HumanHap510S at the Genotyping Shared Resources at the Mayo Clinic in Rochester, MN, which consists of a total of 1,055,048 markers [62, 63]. Raw data were normalized according to the procedure outlined in [64].

172 cell lines (60 Caucasian, 53 African-American, 59 Han Chinese-American) had both gemcitabine cytotoxicity measurements and genome-wide marker intensity data. To illustrate the application of the kernel permutation approach, we selected

one chromosome (chromosome 3) from the genome-wide data. To control for the possibility of population stratification, which can lead to spurious associations, we used the method developed by [65], which uses a principal components analysis (PCA) to adjust for stratification. At each marker, a linear regression model was fit with PCA-adjusted IC50 as the outcome and intensity at that marker as the explanatory variable; these models produce the marker-level tests.

We analyzed these data using the kernel-based approach described in Section 2.2 with a bandwidth of 50 markers and the log transformation. The results are shown in Figure 3.2. Note the presence of a peak at 102.6 Mb; this genomic region was also illustrated in Figure 1.1. The red line indicates the FWER-controlled, chromosome-wide significance threshold at the $\alpha = 0.1$ level. As the figure indicates, there is insufficient evidence in this study to establish a CNV association involving response to gemcitabine ($p = 0.16$) after controlling the chromosome-wide FWER. Other choices of bandwidth and transformation produce qualitatively similar, although somewhat less significant, results.

Copy number variation in the region of chromosome 3 at 102.6 Mb, which is in close proximity to the gene ZPLD1, has been found by [66] to be associated with childhood obesity. An earlier analysis of this data by [35] indicated suggestive evidence that this region harbors a CNV association with gemcitabine response but lacked a formal way to control the error rate at the chromosome-wide level. This example illustrates the need for the more rigorous approach we develop here. The lack of significance in this example is perhaps not surprising, in that 172 subjects is a relatively small sample size for a CNV association study.

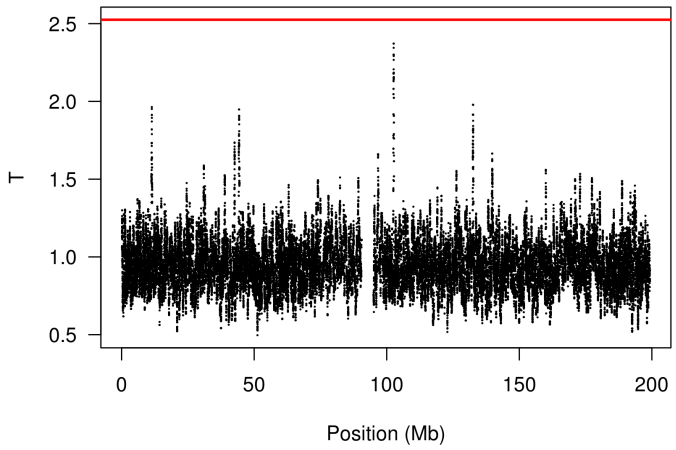

Figure 3.2: Analysis of the gemcitabine data (Chromosome 3) using the proposed kernel aggregation method. The kernel aggregations T_j are plotted against chromosomal position. The red line indicates the cutoff for chromosome-wide FWER significance at the $\alpha = .1$ level.

3.4 Simulation

3.4.1 Spike-in data design

As illustrated, the permutation approach is a valid way to assess the significance of the proposed kernel-based CNV-phenotype association test. In order to study the power of the proposed approach to detect CNV-phenotype associations, we simulate CNVs and their corresponding intensity measurements, LRR. The validity and accuracy of our conclusions rely on how realistic the simulated data is, so we need to put careful thought into simulating this data in as realistic a manner as possible. Here in the thesis, I use the spike-in design that is described in [35].

The basic design of our simulations is to use real data from the gemcitabine study described in [35], "spike" a signal into it, then observe the frequency with which we can recover that signal. We fit a circular binary segmentation model [30, 31] to estimate underlying mean intensity for every marker along the chromosome. Then We calculate the residuals by subtracting the actual intensity measurement from the es-

timated mean. These residuals form a matrix representing measurement errors from real data and we use these residuals to simulate our LRR noise. We pick chromosome 3 of the gemcitabine pharmacogenomic study for simulation and this residual matrix, denoted R, has 172 rows (one for each cell line) and 70,542 columns (one for each marker).

Our simulations involve short genomic regions containing 200 markers in which a single CNV is either present or absent. The length of the CNV varies from 10 to 50 markers. We randomly select residuals from the above residual matrix to simulate LRR noise over our study genomic regions. Then add in a signal. Letting i denote subjects and j denote markers, the following variables are generated: z_i, an indicator for the presence or absence of a CNV in individual i; x_{ij}, the intensity measurement at marker j for individual i; and y_i, the phenotype for subject i. We focus here on a random sampling design in which the outcome is continuous. In the random sampling design, the CNV indicator, z_i, is generated from a Bernoulli distribution, where $\gamma = \Pr(z_i = 1)$ is the frequency of the CNV in the population. Meanwhile, $y_i|z_i$ is generated from a normal distribution whose mean depends on z_i.

For each simulated data set of every subject, 200 markers were randomly selected from the columns of R. The measurement error for simulated subject i was then drawn from the observed measurement errors at those markers for a randomly chosen row of R. The random selection of markers would remove the possibility of bias from correlation among the intensities of neighboring markers. Thus, within a simulated data set, all subjects are studied with respect to the same genetic markers, but the markers vary from data set to data set. Simulating the data in this way results in all the features of outliers, heavy-tailed distributions, skewness, unequal variability

among markers, and unequal variability among subjects that are present in real data.

The intensity measurements $\{x_{ij}\}$,as mentioned, derive from these randomly observed residuals in the real data. To the noise, we add a signal(mean structure) that depends on the presence of the simulated CNV, z_i. The added signal is equal to zero unless the simulated CNV region; otherwise the added signal is equal to the standard deviation of the measurement error times the signal to noise ratio. Thus, adding the signal value to independent measurement errors, we generate our simulated intensities. Choosing an appropriate signal to noise ratio is less obvious. For the amount of noise, the standard deviation of the residual values is 0.9. Signal, however, depends on a number of unknown and poorly understood factors. We would simulate CNVs with a signal of mean shift 0.72 and thus employed a signal to noise ratio of 0.8 for simulation, which corresponded roughly to a medium-sized detectable signal based on our inspection of the gemcitabine data. In such a construction, phenotype and intensity measurement are conditionally independent given the latent copy-number status z_i. An illustration of the spike-in process is given in Figure 3.3.

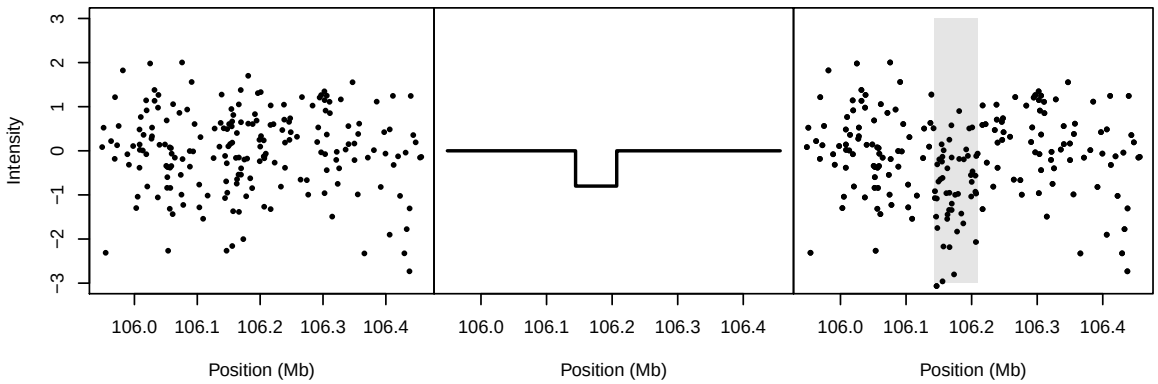

Figure 3.3: Illustration of spike-in simulation design. *Left:* The noise, randomly drawn from among the estimated measurement errors for a single cell line. *Middle:* The spiked-in signal. *Right:* The resulting simulated data.

Using this procedure, the simulated data appears similar to the real data. For the Illumina Human1M-Duo BeadChip, which has a median spacing of 1.5 kb between markers, 200 markers corresponds to simulating a 300 kb genomic region. We varied the length of the CNV from 10 to 50 markers, corresponding to a size range of 15 to 75 kb. For the simulations presented in the remainder of this section, we used a sample size of $n = 1,000$ and an effect size (mean divided by standard deviation) of 0.4 for the continuous outcome. All association tests are conducted with type I error rate of 0.05.

Using such simulation data, we compare the power while varying CNV sizes, transformations of p-values or kernels. For each setting, 1000 independent data sets are generated and analyzed. Power is defined as the fraction of data sets in which CNV-phenotype association is declared. The association test at each marker would derive from a linear regression model between intensity and phenotype. One would prefer a method that not only detects CNV associations but also identifies their boundaries. Here I focus only on detection of phenotype-associated CNV in my thesis.

3.4.2 Comparison of transformation

For the various transformations proposed in Sections 2.2.2 and 2.2.3 for different association tests, we evaluate the relative impacts of transformation and association direction on power. First, we here set CNV frequency to be 10%. In order to isolate the effect of transformation, we set bandwidth of the kernel to be the "optimal bandwidth", which is chosen to match the number of markers for the underlying CNV and thus results in the maximum power to detect a CNV-phenotype association. In practice this approach is not feasible since the size of the underlying CNV is unknown.

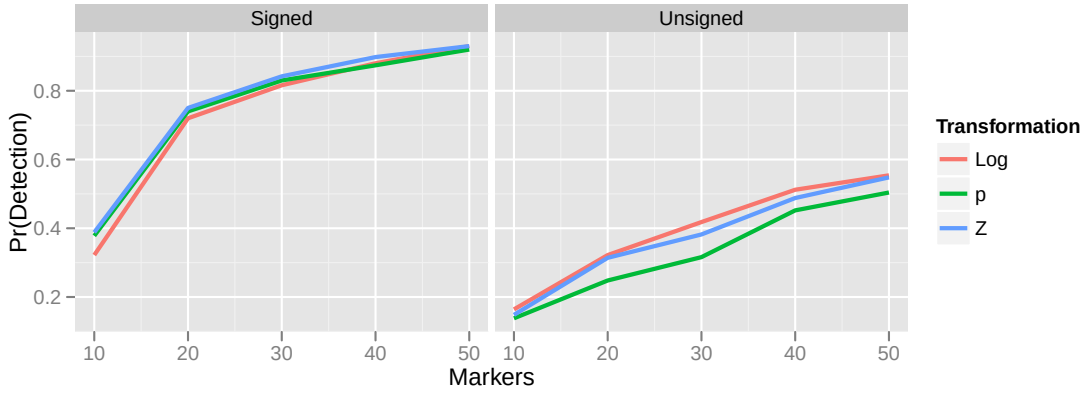

Figure 3.4: Effect of transformation choice and direction of association on power. Population CNV frequency was set to 10%; optimal bandwidths used.

The results in Figure 3.4 demonstrate the impact of transformation choice on power. The figure illustrates a basic trend that held consistently over many CNV frequencies and bandwidth choices. Various choices of transformation produce consistent results for both association test settings by comparing the left and right halves of the figure. Furthermore, if the direction is available along with the test, all transformations by incorporating the direction of association perform much better than ignoring the direction of association with regard to power. Besides, although various transformations do not alter power dramatically, the normalizing transformation (Z) is the most powerful for signed test results, while the log transformation obtains the highest power for unsigned tests. In the results that follow, unless otherwise specified, I employ the normalizing transformation for signed test results and the log transformation for unsigned tests.

3.4.3 Comparison of kernel choice

As illustrated in Section 2.2.1, there are two important factors with regards to kernel. Thus here in this section, we examine two aspects of kernel choice: bandwidth imple-

mentation (constant-width vs. constant-marker) and kernel shape (flat vs. Epanechnikov). When all markers are equally spaced, the constant-width and constant-marker kernels are equivalent. To evaluate the impact of bandwidth on power when markers are unequally spaced, we selected at random a 200-marker sequence from chromosome 3 of the Illumina HumanHap 550K genotyping chip and spiked in CNVs of various sizes. We specified five bandwidth corresponding to window size varying from 10 to 50 markers. The optimal bandwidth (either in terms of the number of markers or base pairs spanned by the underlying CNV) was chosen for each method.

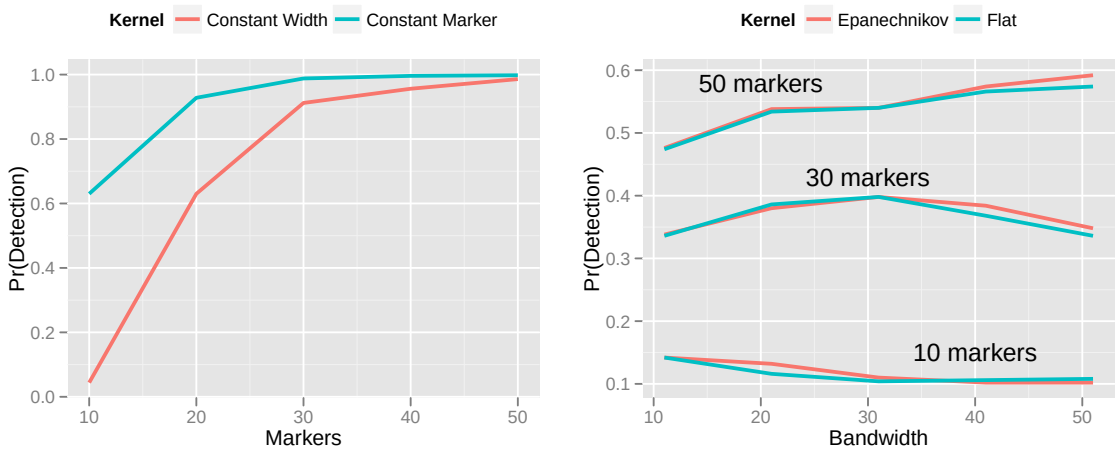

Figure 3.5: Effect of kernel choice on power. *Left:* Constant-width kernel vs. constant-marker kernel. *Right:* Flat vs. Epanechnikov kernel. In both plots, population CNV frequency was 10%, test results were unsigned, and the log transformation was used.

The left side of Figure 3.5 presents the results of this simulation comparing kernels with two different bandwidth implementations. The constant-marker approach is substantially more powerful through all bandwidth settings. When the number of markers within each window is not held constant, the aggregation measure T_j is more highly variable for some values of j than others. This causes the null distribution of $T_{\max}$ to have thicker tails, which in turn increases the p-value for the observed $T_{\max}$,

thus lowering power. This phenomenon manifests itself most dramatically for small bandwidths. Consequently, throughout the rest of this chapter, we employ constant-marker kernels for all analyses.

The right side of Figure 3.5 presents the results of comparing the kernel shape between the flat kernel described in (2.2) and the Epanechnikov kernel described in (2.3). We make several observations: (1) The shape of the kernel has only a limited effect on power; the performance of two different kernel functions with regard to power seems similar. (2) The kernel approach is relatively robust to choice of bandwidth; even 5-fold differences between the test bandwidth and optimal bandwidth do not dramatically reduce power. (3) Nevertheless, the optimal bandwidth does indeed perform best when the number of markers included in the kernel matches the true number of markers spanned by the CNV. A larger window may include too many non-informative markers and lead to a reduction in testing power; a smaller window may ignore informative markers and thus decrease the power. Thus, choosing different bandwidth would change the power. (4) The Epanechnikov kernel is slightly more robust to choice of bandwidth than the flat kernel is. This makes sense, as the Epanechnikov kernel gives less weight to the periphery of the kernel.

3.4.4 Comparison of kernel-based aggregation and variant-level testing

Lastly, we compare the kernel-based aggregation approach with variant-level testing. To implement variant-level testing, each sample was assigned a group ("variant present" or "variant absent") on the basis of whether a CNV was detected by CBS. A two-sample t-test was then carried out to test for association of the CNV with the phenotype. This variant-level approach was compared with kernel-based aggregation

of marker-level testing for a variety of bandwidths. The results are presented in Figure 3.6.

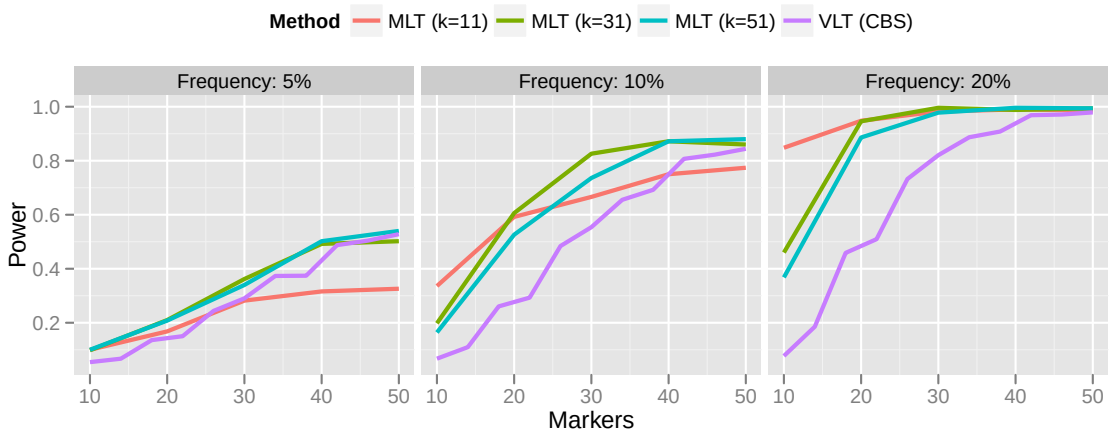

Figure 3.6: Power comparison of variant-level testing (using CBS for CNV calling) with marker-level testing (using kernel-based aggregation).

For rare CNVs (5% population frequency), the power of the variant-level approach and the aggregated marker-level approach are comparable. However, for more common CNVs, the marker-level approach offers a substantial increase in power. For the most part, this increase in power persists even when the bandwidth is misspecified. Only when the bandwidth was too small (selecting a 10-marker bandwidth for a 50-marker CNV) did the variant-level approach surpass marker-level aggregation.

Generally speaking, these results are consistent with the findings reported in [35], who found that variant-level tests have optimal power relative to marker-level tests when CNVs are large and rare; conversely, marker-level tests have optimal power relative to variant-level tests when CNVs are small and common. This is understandable given the limited accuracy of calling algorithms for small CNVs.

Comparing the results in Figure 3.6 with the results of Breheny et al. [35], who aggregated marker-level tests by applying CBS to the p-values as described in Section 3.2.1, we find that the kernel approach is a substantially more powerful method for aggregating marker-level tests than a change-point approach. Specifically, Breheny *et al.* found that the change-point approach had very low power at 5% frequency – much lower than the variant-level approach. On the other hand, in the same setting we find that the kernel approach is comparable to, and even slightly more powerful than, the variant-level approach. Furthermore, as discussed in Section 3.2.1, a change-point analysis of marker-level tests also relies on exchangeability, which does not always hold. Thus, the methods developed here in the thesis are both more powerful and achieve better control over the FWER than the change-point analysis described in [35].

A potential drawback of the kernel approach is the need to specify a bandwidth. This makes the robustness of the method to bandwidth misspecification, as illustrated in Figure 3.6, particularly important because in practice it is difficult to correctly specify the bandwidth *a priori*. Indeed, it is possible that multiple CNVs associated with the outcome are present on the same chromosome and have different lengths. A method that is not robust to bandwidth would be incapable of detecting CNVs. Generally speaking, a bandwidth of roughly 30 markers seems to provide good power over the range of CNV sizes that we investigate here.

Chapter 4 Correlation Method and Its Family Wise Error Rate Control

4.1 Introduction

I introduced kernel-based aggregation method to determine the CNV-phenotype association. As p-values are correlated with each other among markers, it is complicated to estimate the exact null distribution when applying the method. There is no specific computational approaches developed to estimate the null distribution when applying the kernel-based method. I applied a permutation procedure to our kernel-based framework in Chapter 3. I demonstrated that this method provides evidence of association between phenotype and genotype while preserving accurate FWER. However, the permutation approach has its own limitations as mentioned earlier. One important drawback for analysis is the computational burden of the method. For simple tests such as the linear regression tests we used in the gemcitabine study, the burden is quite manageable. On our machine (using an Intel Xeon 3.6 GHz processor), it takes under a second to perform the 70,542 marker-level tests on chromosome 3 and under 0.1 seconds to perform the kernel aggregation. Carrying out 1,000 permutation tests took 1,000 times longer: 15 minutes to carry out all the permutation tests and 21 seconds to perform all the kernel aggregation. Extrapolating a genome-wide analysis would take 3.5 hours. These calculations, however, are for simple marker-level tests and a fairly small sample size ($n = 172$). Larger studies will increase the computation burden linearly (*i.e.*, doubling the subjects should double the computing time), but more complicated marker-level tests based on nonlinear, mixed-effects, or mixture models would require substantially more time. But it is worth pointing out that kernel aggregation itself does not consume time but the estimation of the null distribution. As seen in Figure 3.2, the black dots may be calculated rapidly; the red line is what requires the permutation testing. It is worth further research to discover

ways to speed up the approach.

Since p-values are statistically dependent, the exact null distribution relies on the correlation structure of p-values. In this chapter, a correlation-based approach with a model-based formulation that avoids the need for permutation testing was considered to apply to our kernel-based method in order to speed up the analysis. I present and demonstrate the procedure in details to approximate the joint distribution of the test statistics. I show through simulations that this approach not only provides an accurate error control, but also is much faster than the permutation approach.

4.2 Basic idea about correlation method

4.2.1 Correlation approach

Applying kernel-based aggregation of marker-level association test, we consider a two-stage procedure. From association testings for every marker, we obtain a set of test results $\{p_j\}$. Then, when considering quantifying whether or not the data represent compelling evidence for a CNV-phenotype association we use the statistic

$$T_{\max} = \max_j \{T_j\}, \tag{4.1}$$

where T_j is the kernel-based aggregation of p-values in a window. If the tests are directional, with results $\{p_j, s_j\}$, we use $T_{\max} = \max_j\{|T_j|\}$.

To obtain the null distribution of $T_{\max}$, we present an alternative to the permutation approach, which is a correlation-based procedure. We now formally present this procedure to fit our kernel-based aggregation framework as follows and show that it preserves the family-wise error rate for the problem of CNV association testing.

Step 1: Estimate a correlation matrix Σ that defines the relationship among z-statistics under null. First of all, we assume that the correlation completely describes the dependence between the series of z-statistics. If Σ is positive definite, then the Cholesky decomposition, $\Sigma = CC^T$, can be obtained. In many situations, the correlation of the series of original p-values is complicated to estimate from real data. I will later give details of calculating $\hat{\Sigma}$ for the CNV-phenotype association study under assumption of simple linear model.

Step 2: Generate vectors of dependent z-values mimicking the original z-values from the real data. Geometrically, a symmetric and positive definite matrix C will transform uncorrelated variables to dependent z-values. According to [40], generate a random vector U from independent uniform (0,1) distribution and let $Z = \Phi^{-1}(1 - U)$ where $\Phi(.)$ was the cumulative distribution function of a standard normal random variable. Given a random vector of i.i.d uncorrelated standard normal variables Z, the Cholesky transformation maps the variables Z into variables CZ with covariance matrix $Var(CZ) = CVar(Z)C^T = CIC^T = \Sigma$. Thus, CZ reflects the distribution of test statistic under null that there is no association between genotype and phenotype.

Step 3: Construction of empirical null distribution of the test statistic. Based on the multivariate normal distribution CZ described in Step 2, we could create series of test statistics from such distribution. If the direction of association is available, we could calculate p-values for both signed and unsigned cases. In this way, we may generate an arbitrary number of vectors of dependent p-values among markers. It is computationally less demanding since it does not involve repeated analysis of simulated datasets. For each vector of dependent p-values for the bth Monte Carlo sample, we calculate the proposed test statistic applying kernel-based method to yield $\{T_{\max}^{(b)}, b = 1, ..., B\}$ for any choice of transformation and kernel in (2.1). Finally, we use the empirical

CDF of those draws to obtain the estimate $\hat{F}_0 = \sum_{1 \le b \le B} I[T_{\max}^{(b)} \le x]/B$.

Step 4: The calculation of the empirical overall p-value. By the estimation of null distribution function in Step 3, we obtain a test for the presence of a CNV-phenotype association based on $p = 1 - \hat{F}_0(T_{\max})$.

4.2.2 Replacing correlation among z statistics with correlation of intensities

It is worthy to point out that this approach depends on the estimation of correlation matrix of z-statistic under null. But it is always complicated to estimate. Suppose our analysis data consist of a set of marker genotypes together with phenotypic trait values measured on each individual $(\boldsymbol{X}^i, y_i)$, $i = 1, 2, ...n$, where $\boldsymbol{X}^i = (X_{i1}, ..., X_{iJ})^T$ is a vector of the copy number intensities for every marker for subject i and y_i is the phenotype for subject i. Assume simple linear regression $y = \beta_0 + \beta_{1j} x_j + \epsilon$ between phenotype y as the outcome and intensity x_j at jth marker as the explanatory variable for every marker, where ϵ has a normal distribution with mean 0 and standard deviation σ. By standardization, we define $x_{ij} - \bar{x_j} = x'_{ij} sd_{x_{ij}}$ and $y_i - \bar{y} = y'_i sd_{y_i}$, where we denote standardized x'_{ij} and y'_i.

Under the simple linear regression model, we have the z-statistics under H_0:

$$\begin{aligned} Z_j &= \frac{\hat{\beta_{1j}} - 0}{s.e.} \\ &= \frac{\frac{\sum (x_{ij} - \bar{x_j})(y_i - \bar{y})}{\sum (x_{ij} - \bar{x_j})^2}}{\hat{\sigma} \sqrt{1 / \sum (x_{ij} - \bar{x_j})^2}} \\ &= \frac{\sum sd_{x_{ij}} x'_{ij} sd_{y_i} y'_i}{\hat{\sigma}(n-1) sd_{x_{ij}}^2} \sqrt{(n-1) sd_{x_{ij}}^2} \\ &= \frac{\sum x'_{ij} y'_i}{\sqrt{(n-1)}}, \end{aligned}$$

where $\sum(x_{ij} - \bar{x_j})^2 = (n-1)sd^2_{x_{ij}}$ and also $\hat{\sigma} = sd_{y_i}$.

Now, correlation between z statistics of any two markers is simplified and we get

$$\begin{aligned} cor(z_j, z_k) &= cor(\frac{\sum x'_{ij}y'_i}{\sqrt{(n-1)}}, \frac{\sum x'_{ik}y'_i}{\sqrt{(n-1)}}) \\ &= cor(\sum x'_{ij}y'_i, \sum x'_{ik}y'_i) \\ &= cor(\sum x'_{ij}, \sum x'_{ik}), \end{aligned}$$

since under the null hypothesis that x_{ij} and y_i are independent.

Therefore,

$$\begin{aligned} cor(z_j, z_k) &= cor(\sum x'_{ij}, \sum x'_{ik}) \\ &= cor(\mathbf{x}_j, \mathbf{x}_k) \end{aligned}$$

We summarize and present the following theorem.

Theorem 2. *Assume a simple linear regression model was fit at each marker between phenotype and genotype. Let H_{0j} denote the hypothesis that the phenotype, y_i, and the CNV intensity at marker j, X_{ij}, are independent. Then, under the global null hypothesis H_0 that y_i is jointly independent of $\{X_{ij}\}$, correlation structure among z statistics under H_0 exactly equals the correlation matrix of the intensities among markers under assumption of simple linear model.*

By Theorem 2, the correlation structure among z-values under H_0 is the same as the correlation matrix of the intensities under assumption of simple linear model between intensities and phenotypes for every marker. It is worth to note that the

estimate of correlation structure might depend on the knowledge of model assumption between intensities and phenotype. Then we need the fact that correlation is approximately invariant under monotone transformations [40]:

$$Cor(g(X_i), g(X_j)) \approx \frac{[g'(\mu)]^2 Cov(X_i, X_j)}{\sqrt{[g'(\mu)]^4 Var(X_i) Var(X_j)}} = Cor(X_i, X_j) \tag{4.2}$$

For some modelings like logistic regression, the z-value is a monotone function of z-value of linear regression. Applying (4.2), their correlation should be more or less equivalent. Thus, we could extend to some other model assumption between phenotype and genotype by (4.2). If the z-values for some model assumption is highly monotone in comparison with linear regression, we could get approximately the same correlation estimate as the estimate under linear modeling.

4.2.3 Estimate of correlation matrix of intensities among markers

Theorem 2 provides a good way to estimate the correlation structure of z-statistic. Given the structure of intensities among markers, we could use the sample correlation of intensities to estimate the correlation matrix of z-statistic under null hypothesis of no CNV-phenotype association. But estimation of correlation matrix and application of such approach would be easy to achieve and would be well-applied when the number of features is small. When the number of markers is large, it would cause problems. First of all, the high dimensions $J \times J$ matrix calculation will bring statistically efficiency and computational problem. It always increases the computation burden and might run out of memory for high dimensions. Besides, the algorithm in section 4.2.1 requires Cholesky decomposition, which gets slower for high dimension matrix. Hence, estimation of large-scale covariance matrix from real genomic data become an ubiquitous problem. There are a lot of methods proposed in the literatures for estimation under such case, including the spectral decomposition, Bayesian meth-

ods, modeling the matrix-logarithm, nonparametric smoothing, and banding/thresholding techniques [67–73].

It would be important to extend our research to high dimension, which is the typical situation in real data. Therefore, I develop different approaches to estimate the restricted correlation matrix for computation convenience in different circumstances for high dimensions of features. I will talk about it in the following sections.

4.3 Extending correlation approach to "small n, large J" setting

For genomics and transcriptome analysis, estimation of large-scale covariance or correlation matrices is a common problem. From a microarray experiment or CNV data, J markers are being analyzed with J perhaps in the order of 1,000 to 10,000 and thus a correlation of size $J \times J$ has to be calculated. For analysis involving real data, we do not know where CNVs are located before analysis. We normally collect more and more intensity data that might include the possible CNV to detect the CNV-phenotype association and detect the CNV location. Hence, it is pretty common to involve thousands or millions of markers from part of or the whole chromosome. Meanwhile, it often encounters with a limited number of samples n. A common key problem for all such data is that how we should obtain an accurate and reliable estimate of the population covariance matrix when a data set includes a large number of variables but only contains comparatively few samples ($n << J$). Under such "small n, large J" setting, singular value decomposition (SVD) and principal component analysis (PCA)are valuable tools and common techniques for analysis of such multivariate data.

Principal component analysis (PCA) (Jolliffe 1986) is a popular data-processing and dimension-reduction technique for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information. It is a widely used mathematical tool for high dimension data analysis. PCA provides a guideline for how to reduce a complex data set to one of lower dimensionality to reveal any hidden, simplified structures that may underlie it. It extracts the most important information in such a way as to highlight their similarities and differences from the data table and compress the size of the data set by keeping only this important information. Then it can simplify the description of the data set and analyze the structure of the observations and the variables. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information. Overall, PCA is one of the simplest and most robust ways capable of reducing dimensions and revealing relationships among data items. It has been applied in numerous fields such as engineering, biology, and social science. Some interesting examples include handwritten zip code classification (Hastie,Tibshirani, and Friedman 2001) and human face recognition (Hancock, Burton, and Bruce 1996). Recently PCA has been used in gene expression data analysis (Alter, Brown, and Botstein 2000). Hastie et al. (2000) proposed the so-called gene shaving techniques using PCA to cluster highly variable and coherent genes in microarray datasets. Strictly speaking, singular value decomposition is a matrix algebra trick which is used in the most common algorithm for PCA. PCA can be computed via the singular value decomposition(SVD) of the data matrix.

According to Theorem 2, correlation structure among z-statistics Σ under H_0 exactly equals the correlation matrix of the intensities among markers. Let X be $n \times J$ original intensity matrix, where n is the sample size and J is the number of

total markers. Applying SVD method on it, the singular value decomposition of X is the factorization of X into the product of three matrices,

$$X = UDV^T, \tag{4.3}$$

where U is $n \times n$ orthogonal matrix with $U^TU = I$, V is $J \times J$ orthogonal matrix with $V^TV = I$. We could partition $U = (u_1, ..., u_n)$ and $V = (v_1, ..., v_J)$, where u_j denote the jth left singular vector and v_j denote the jth right singular vector. D is $n \times J$ diagonal matrix with diagonal entries $(D_{11} = \sigma_1) \geq (D_{22} = \sigma_2) \geq ... \geq (D_{rr} = \sigma_r) \geq 0$. Here $\sigma_1, \sigma_2, ..., \sigma_r$ are called the singular values. The singular values are sorted from high to low, with the highest singular value in the upper left index of the matrix. If $rank(X) = r$, then it equals to the number of non-zero singular values on the diagonal. Also, the column space of X is spanned by the first r columns of U and the last $n - r$ columns of U span the null space of X^T. Meanwhile, the row space of X is spanned by the first r columns of V and the null space of X is spanned by the last $J - r$ columns of V.

$$X^TX = (UDV^T)^TUDV^T = VDU^TUDV^T = VD^2V^T, \tag{4.4}$$

$$XX^T = UDV^T(UDV^T)^T = UDV^TVDU^T = UD^2U^T, \tag{4.5}$$

where X^T is the conjugate transpose of X. The right hand sides of these relations describe the eigenvalue decompositions of the left hand sides. Consequently, the squares of the non-zero singular values of X are equal to the nonzero eigenvalues of either XX^T or X^TX. Thus XX^T which is $n \times n$ and X^TX which is $J \times J$ will share n eigenvalues when $n < J$ and the remaining $J - n$ eigenvalues of X^TX will be zero. The columns of V, which are called right singular vectors, are the eigenvectors of X^TX and the columns of U ,called left singular vectors, are eigenvectors of XX^T.

Either the singular vectors of the singular values (or both) are called principle components. The matrix D does not have to be square. Note that for a square, symmetric matrix X, singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem. SVD decomposes a matrix into a set of rotation and scale matrices, which is used in computing the pseudoinverse, matrix approximation, and determining the rank, range and null space of a matrix.

Let $D_r = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix}$, and hence (4.3) will become

$$X = \begin{pmatrix} U_r & U_{n-r} \end{pmatrix} \begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_r^T \\ V_{J-r}^T \end{pmatrix} = U_r D_r V_r^T, \tag{4.6}$$

using $U_r = \begin{pmatrix} u_1 & \cdots & u_r \end{pmatrix}$ and $V_r = \begin{pmatrix} v_1 & \cdots & v_r \end{pmatrix}$. This thin SVD is equivalent to the ordinary SVD. Applying (4.6), any $n \times J$ real matrix equivalent to the product of a column-orthogonal $n \times r$ matrix, a diagonal $r \times r$ matrix where the elements sorted in descending order, and a column-orthonormal $J \times r$ matrix. It is a useful result since we could reduce the dimension of the ordinary large matrix into product of simplified matrix. In "small n, large J" setting, the number of the non-zero singular values of such a large matrix X, which equals the rank of X, would be smaller or equals to $min(n, J) = n$. If we know the rank of the intensity matrix to be r, we could get the thin SVD of X. Thus, applying such procedure, we could deal with smaller matrix with the optimal low rank instead of the large matrix.

For calculating the covariance matrix of intensities, there is a direct relation between PCA and SVD in the case where principal components are calculated from the covariance matrix. For PCA to work properly, the first step in PCA is to move the

origin to mean of the data. In our case, we could achieve the first step by finding means of every marker by averaging the columns of X. We then subtract the mean intensity from each intensity of the data set (ie each row of X) to create the mean centered data vector. This produces a data set whose mean is zero for every marker. It is very easy to compute the covariance matrix from the mean centered data matrix. Furthermore, we apply the method to the correlation rather than the covariances. For correlation matrix, the off-diagonal elements are on the same scale. Also, the correlations derived from covariance estimator are independent of scale and location transformations of the underlying data matrix. Due to such distinct advantages, it is better to work on correlation matrix, which we denote Σ. To achieve the correlation matrix, we need to standardize the original data matrix. For every element of the data matrix x_{ij}, we apply the standardization.

$$x'_{ij} = \frac{x_{ij} - \bar{X}_j}{s_{X_j}}, \tag{4.7}$$

where $\bar{X}_j$ and s_{X_j} are mean and sample standard deviation of jth column intensities for jth marker. For the standardized data matrix which we denote as $\boldsymbol{X}$, we could get the sample correlation matrix.

$$\Sigma = \frac{1}{n-1}(\boldsymbol{X}^T\boldsymbol{X}) \tag{4.8}$$

Applying equation (4.6), here the sample correlation matrix would be

$$\Sigma = \frac{1}{n-1}(\boldsymbol{X}^T\boldsymbol{X}) = \frac{1}{n-1}(VD^2V^T) = \frac{1}{n-1}(V_rD_r^2V_r^T) \tag{4.9}$$

The diagonalization of $\boldsymbol{X}^T\boldsymbol{X}$ yields V, which also yields the principal components of correlation matrix. So, the right singular vectors v_k are the same as the principal components of correlation matrix. Using the decomposition above, we can identify the eigenvectors and eigenvalues for $\boldsymbol{X}^T\boldsymbol{X}$ as the columns of V and the squared diagonal elements of D, respectively. Clearly, the eigenvalues and eigenvectors of correlation

matrix is the same as the eigenvalues and eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. Meanwhile, the latter shows that the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ must be non-negative. Therefore, the sample correlation matrix here is positive semi-definite.

It is a common setting to have much more markers than samples, $n \ll J$. The $J \times J$ correlation matrix itself is therefore very unpleasant to work with because it is very large under "small n, large J" setting. However, by Theorem 2, it suffices to decompose a smaller matrix. In such setting, it is known that $\boldsymbol{X}\boldsymbol{X}^T$ which is $n \times n$ and $\boldsymbol{X}^T\boldsymbol{X}$ which is $J \times J$ will share n eigenvalues and the remaining $J - n$ eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ are all zeros. Therefore, for the eigenvalues which are zeros, we can essentially discard those eigenvalues and the corresponding eigenvectors, hence reducing the dimensionality of the new basis. Instead of dealing with $J \times J$ matrix for Σ, we could focus on the smaller matrix with dimensions $n \times n$ for correlation matrix without losing information since the other $J - n$ dimensions don't contain any additional information. It will be an amazing improvement in speed of computing the estimated correlation matrix. For example, for a $174 \times 70,000$ intensity matrix, we don't need to work with $70,000 \times 70,000$ correlation matrix, which we might not be able to store in our computers. Instead, by the decomposition, the 174×174 matrix, which is a lot faster to obtain, is enough to estimate the correlation matrix. It's really a great result that the estimation of correlation matrix has nothing to do with the large J dimensions but instead only depends on sample size for the matrix dimension. Thus, the smaller n is, the much faster to estimate the correlation matrix of intensities. For standardized intensities $\boldsymbol{X}$, the correlation matrix is equivalent to the variance-covariance matrix. It is worthy to notice that the rank of the variance-covariance matrix is n while the correlation matrix only has a rank of $n - 1$ since it is restricted to have 1's along the diagonal. Therefore, following the above equation (4.9), we obtain the estimated correlation matrix $\frac{1}{n-1}(VD^2V^T)$ and it is equivalent to $\frac{1}{n-1}V_rD_r^2V_r^T$ in a much smaller

dimensions, where $r = n-1$. we only need to know the first $n-1$ left singular vectors. Meanwhile, the decomposition has nothing to do with the right singular vectors u_k. Thus, we could reduce the dimensions a lot and only select $n-1$ vectors for calculation of the correlation matrix. Obviously, the correlation matrix is obtained in a much faster speed. As here the correlation matrix is positive semi-definite, we can denote $\Sigma = CC^T$. By decomposition,$C = \frac{1}{sqrt(n-1)}(VDV^T)$. Thus, given a random vector of i.i.d uncorrelated standard normal variables Z, variables CZ follows a multivariate normal distribution with covariance matrix $Var(CZ) = CVar(Z)C^T = CIC^T = \Sigma$.

This method works especially well for "small n, large J" setting. The small n makes dimension reduction for calculating correlation matrix and hence speed it up a lot compared to directly estimating the large correlation matrix. The comparisons of performance will be shown in the later sections.

4.4 Extending correlation approach to "large n, large J" setting

There is a computation burden problem to estimate the high-dimensions correlation matrix and use such large correlation matrix to generate the multivariate normal random variables under null hypothesis. The time consumption increases as the dimension increases. I have proposed SVD decomposition method for the case of "small n, large J", which leads to the dimension reduction. Thus, we only need to deal with low-dimensional matrix with dimension depending on the small n to estimate the correlation matrix. However, this method is insufficient when n gets large. Instead, we hope to impose a low-dimensional structure on the estimator and aim to get a sparse correlation matrix to decrease the computation burden. Such sparse correlation matrix gets rid of part of correlation information by setting them to be zero.

Hence it would not be positive-definite any more. We have to additionally shrinkage this sparse matrix to be positive-definite for further generation of multivariate normal vectors for "large n, large J" setting. Actually, it is pretty difficulty to get a sparse and positive-definite estimated correlation matrix to accurately estimate the population correlation matrix among genes. I tried a couple of ways and luckily find out an appropriate method to estimate the true correlation matrix under "large n, large J" setting.

Biologically, the correlation between any two markers is a decreasing function of distance. In other word, two nearby markers would be more highly correlated than two markers far apart. If two markers are far enough apart, we may ignore their correlation by considering it to be zero. Setting those small correlations to be zeros, we could restrict the correlation matrix to be a banded correlation matrix instead of calculating every correlation entry. Such a sparse correlation matrix would improve the efficiency for computation under high dimensions of markers. On one hand, there is a storage format for sparse matrix that stores only the nonzero entries in column order and hence avoid the storage of the zero entries. Thus it requires much less memory to store the matrix. Moreover, the computation focuses on the numerical values of the nonzero entries so as to reduce memory usage and avoid unnecessary numerical operations. In this way, calculations based on sparse matrixes would provide a significantly higher calculation efficiency. What's important, it is worthy to notice that such banded sparse correlation matrix contains most of important correlation entries without losing much information.

Here Figure 4.1 is a filled contour plot of sample correlation matrix under null for 200 markers with the middle CNV size=30 for 10000 subjects and signal-to-noise

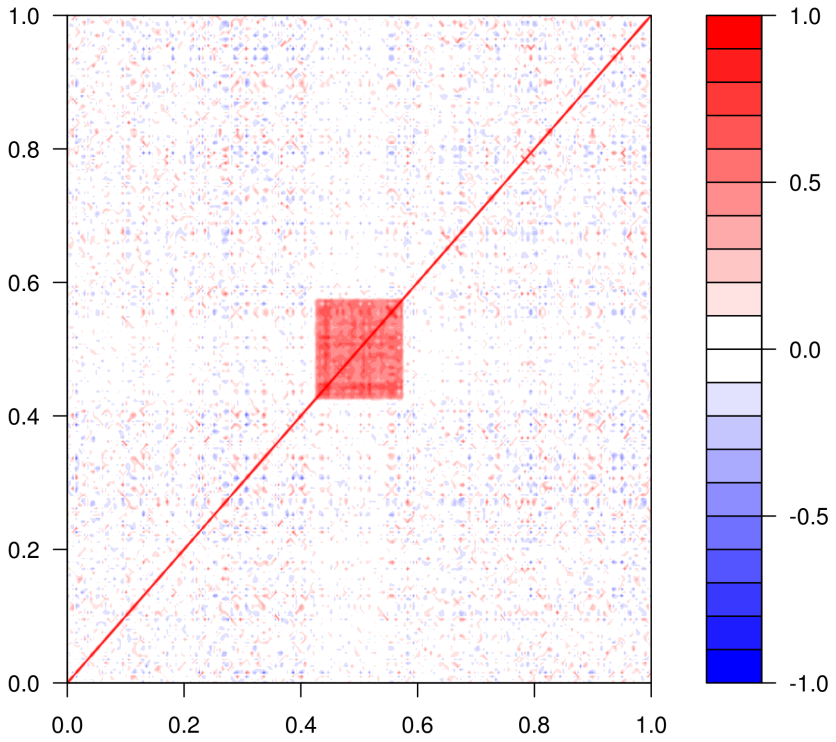

Figure 4.1: This is the sample correlation matrix. Red values are high correlation; the only legitimate correlations are located in the CNV. The rest is just noise.

ratio=2.

In this plot, the red square with high correlations focus on the middle CNV markers while noise smaller correlations are surrounding for the rest markers far apart. Hence, it would be applicable and reasonable to ignore those noise and set those correlations to be zeros. Instead, we could just give an estimate of middle banded correlations for the neighboring markers as the estimate of a huge correlation matrix. Including the whole red part, we would not lose much information and get the estimate more accurate and close to the true correlation. Define d to be the number

of diagonals we calculate the pairwise correlation for the sparse matrix. d choosing would effect the estimate of banded matrix. If using the true CNV size as d, the estimated sparse correlation matrix would be the following plot Figure 4.2 and it contains the whole central correlation of the sample correlation.

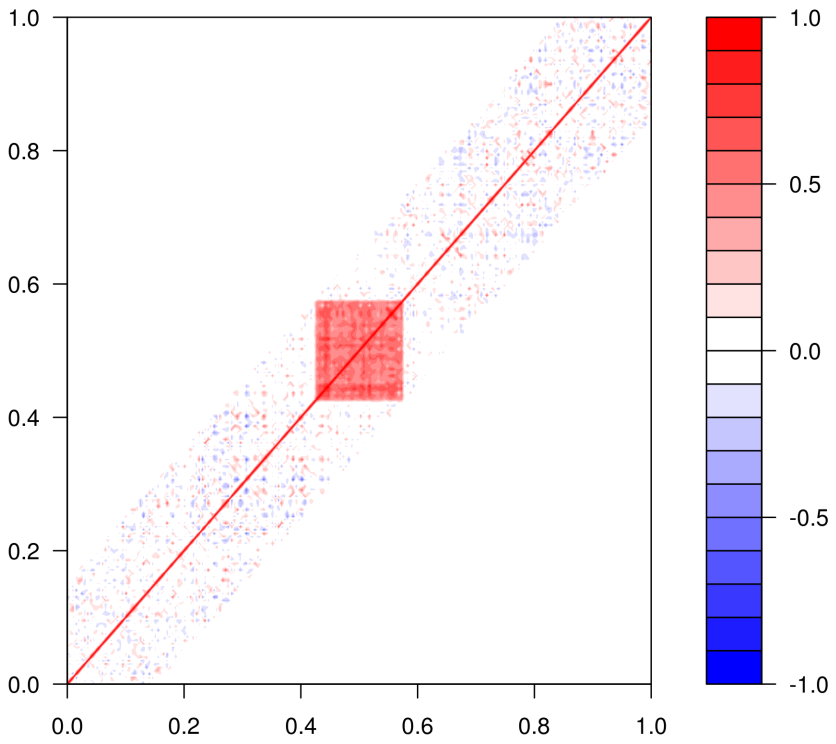

Figure 4.2: This is sparse matrix with $d = 30$ on both sides. It has the same central banded correlation with the rest correlations equal to zero.

But if choosing a different d, the estimate would be very different. As seen in Figure 4.3, the estimate loses a lot of important correlations when choosing half of the true CNV size.

From the above demonstration, we must keep the central main correlation part to

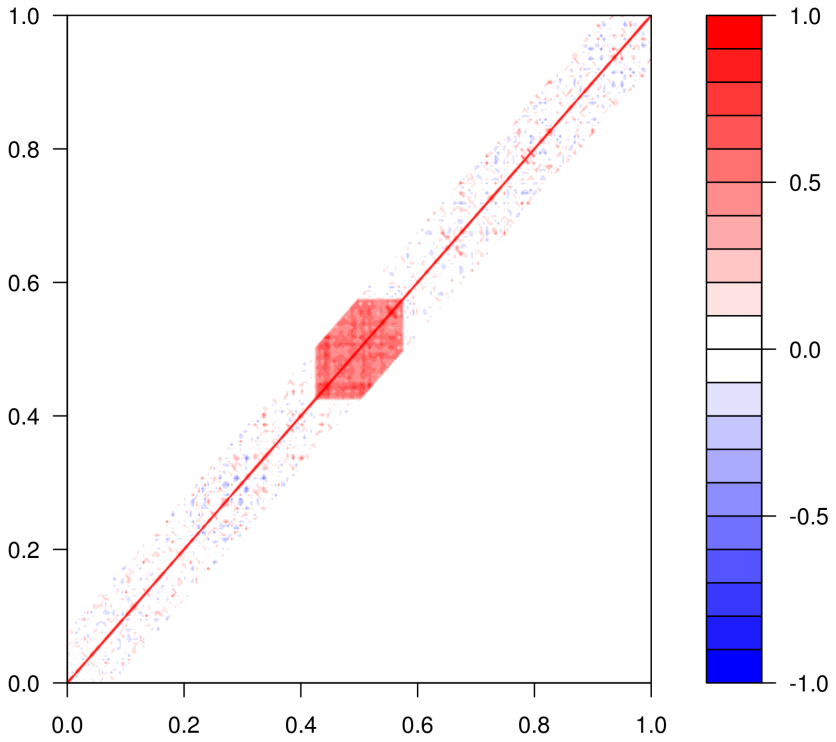

Figure 4.3: It is also a sparse estimate of correlation matrix with $d = 15$ on each side. The important part of the correlation was chopped off.

estimate the true correlation matrix accurately. To keep the central high correlation part of the sample correlation matrix, it would be best to choose d at least equal to the CNV size for the sparse matrix. Otherwise, the estimation would lose a lot of important correlation information and hence gets far away from the whole sample correlation.

However, this sparse estimate would cause singular problems. Such sparse banded correlation matrix might not be positive definite any more. A naive strategy to obtain a positive definite estimator of covariance proposed by Rebonato (1999) runs

as follows: take the sample covariance and decompose the covariance matrix into its eigenvectors and eigenvalues, set the negative eigenvalues to 0 or $(0 + \epsilon)$, and then rebuild the covariance matrix. Higham (2001) uses an optimization procedure to find the nearest correlation matrix that is positive semi-definite. Grubisic and Pietersz (2003) have a geometric method they claim outperforms the Higham technique. Incidentally, some more recent twists on Rebonato's paper are Kercheval (2009) and Rapisardo (2006) who build off of Rebonato with a geometric approach. They also proposed several algorithms for computation to find its nearest correlation matrix. However, such nearest positive definite covariance matrix could not be sparse and thus it is computationally by far demanding for the very large dimensions commonly encountered in genomics problems. Therefore, we aim to keep the estimate of correlation matrix symmetric positive definite and sparse so that it will be a covariance matrix of further multivariate normal distribution and meanwhile reduce the computation burden.

Special care should be taken in the construction of algorithms that create a large symmetric sparse positive definite correlation matrix with the goal of reducing computation time and memory usage. I present shrinkage strategy here to create such sparse positive definite large-scale correlation matrix.

4.4.1 Introduction to shrinkage approach

A fundamental principle of statistical decision theory is that there exists an interior optimum in the trade-off between bias and estimation error. We could shrinkage the unbiased estimator full of estimation error towards a fixed target represented by the biased estimator. Stein(1956) showed that shrinking sample means towards a con-

stant would, under certain circumstances, improve accuracy. In the case of estimated unbiased sample covariance matrix, those estimated coefficients that are extremely high tend to contain a lot of positive error and therefore need to be pulled downwards to compensate for that. Similarly, we compensate for the negative error that tends to be embedded inside extremely low estimated coefficients by pulling them upwards. We call this the shrinkage of the extremes towards the center to increase the accuracy.

Consider the well-known bias-variance decomposition of the mean square error (MSE) for the sample covariance, i.e. $MSE(S) = Bias(S)^2 + Var(S)$. We could obtain an improved covariance estimator is variance reduction. Here we propose "shrinking" or more general "biased estimation" as a means of variance reduction of sample covariance matrix. If properly implemented, this shrinkage would clearly fix the problem of the sample covariance matrix described above to be positive definite and well-conditioned. A recent analytic result was proposed by Ledoit and Wolf (2003) to construct an improved covariance estimator that is not only suitable for small sample size n and large numbers of variables J but meanwhile is also completely inexpensive to compute [74]. They suggested the linear shrinkage approach to combine both single-index covariance matrix estimator and sample covariance matrix estimator in a weighted average. And they select the optimal shrinkage intensity through explicitly minimizing a risk function for example the mean squared error (MSE). It was based on the full sample covariance matrix. As illustrated earlier in our cases, the sparse covariance matrix keeping the main high correlation part is reasonable and accurate to get close to the full sample covariance matrix. Working on such sparse covariance matrix, the key problem is that if we can figure out some similar shrinkage procedure to get it positive definite and keep it sparse in the meanwhile.

Shrinkage here is applied to the correlations rather than the covariances. This has two distinct advantages. First, the off-diagonal elements determining the shrinkage intensity are all on the same scale. Second, the correlations derived from the resulting covariance estimator are independent of scale and location transformations of the underlying data matrix. Working on the sparse matrix, we shrink the off-diagonal entries of the sparse correlation matrix to zero gradually to reach the aim of positive definite matrix. We try the simple shrinkage on each entries of the sparse matrix.

In the following sections, I briefly review the general principles behind shrinkage estimation of the sparse and positive definite correlation matrix and discuss an analytic approach to determine the optimal shrinkage level and the number of sparse diagonals.

4.4.2 Shrinkage estimation of sparse positive definite correlation matrix

As discussed above, we would get a symmetric sparse sample correlation matrix when choosing an appropriate bandwidth. However, such sparse sample correlation matrix could not be positive definite and thus it would have a lot of trouble when creating the multivariate normal random vectors under the null. In this section I suggest using such sparse matrix obtained from the whole sample correlation matrix through a transformation called shrinkage.

Define d to be the number of diagonals we calculate the pairwise correlation. First of all, it is noteworthy that here the d has nothing to do with the bandwidth that we need for kernel aggregation method. d is used to obtain a sparse correlation matrix to estimate the actual correlation while bandwidth is to calculate a finite set of aggrega-

tions $\{T_j\}$ leading to significance testing based on the statistic $T = \max_j\{T_j\}$. From Figure 4.3, it is obvious that the choosing of d would effect the estimation of correlation matrix. An appropriate d would have the sparse correlation matrix close to the sample correlation and otherwise would get the estimation far away from the sample correlation matrix. Thus, the sparse matrix with different d will produce different estimation through shrinkage. I will talk more about the effect of d on estimation of correlation matrix.

In order to obtain positive definite correlation matrix and keep it sparse as well, we have to shrink the elements on the sparse matrix. As we know, the correlation between any two markers is a decreasing function of distance. Two nearby markers would be highly correlated while two markers far apart are less correlated. Thus, it is reasonable to consider shrinking the correlations by distance. We could shrinkage the correlations a little bit for highly correlated marker and meanwhile shrink the correlations more and more as the markers get farther and farther. In some sense, the sparse matrix would be much closer to be positive definite when the correlations are decreasing gradually by distance.

Thus, I consider the exponential decline by distance. Let λ be the shrinkage intensity and j be the number of markers away between two markers. For two markers with j markers away, I shrinkage the correlation by $(1-\lambda)^j$. For any correlation element, the new correlation element $cor(i, i+j)'$ through shrinkage

$$cor(i, i+j)' = cor(i, i+j) * (1-\lambda)^j, \tag{4.10}$$

In other word, working on the sparse matrix with calculated pairwise sample correlations for d diagonals, every element with sample correlation $cor(i, i+j)$ between marker i and marker $i+j$ would transfer into $cor(i, i+j) * (1-\lambda)^j$. It makes some

sense because the correlations of markers far apart would shrinkage a lot more than the nearby markers. After such shrinkage step, we generate a sparse matrix with the gradually decreasing correlations by distance. It will become positive definite when we choose an appropriate shrinkage intensity λ.

4.4.3 Selection of the number of sparse diagonals and appropriate shrinkage intensity

By the shrinkage procedure, we are allowed to construct an improved correlation estimator in a sparse and positive definite matrix that is not only suitable for large sample size n and large numbers of variables J but at the same time is also completely inexpensive to compute. The larger the λ, the more shrinkage the sparse correlation matrix gets. Thus, it will become positive definite more easily but meanwhile it changes the original correlations a lot and gets far away from the original sparse sample correlation. A key question in this procedure is how to select an optimal value for the shrinkage intensity. We wish to get a positive-definite correlation matrix but shrinkage the correlations at minimum level. One common but computationally intensive approach to estimate the minimizing λ is by using cross-validation. Another widely applied route to inferring λ views the shrinkage problem in an empirical Bayes context. In our case here, the optimal shrinkage level could be determined analytically. We can simply choosing minimum λ that guarantees the positive-definite structure of the sparse correlation matrix without the need of specifying any underlying distributions, and without requiring computationally expensive procedures such as MCMC, the bootstrap, or cross-validation. In other word, the optimal shrinkage level λ is determined by how sparse of the correlation matrix.

Setting different number of sparse diagonals d, shrinkage intensity is obtained by choosing minimum λ that guarantees the positive-definite structure of the sparse correlation matrix. The relationship between d and λ is presented in Figure 4.4. It appears that shrinkage intensity λ becomes smaller as the number of sparse diagonals increases. It makes sense in some sense because the sparse correlation matrix gathers more information when d is larger and then it is closer to the sample correlation which is positive semi-definite and hence it does not require much shrinkage.

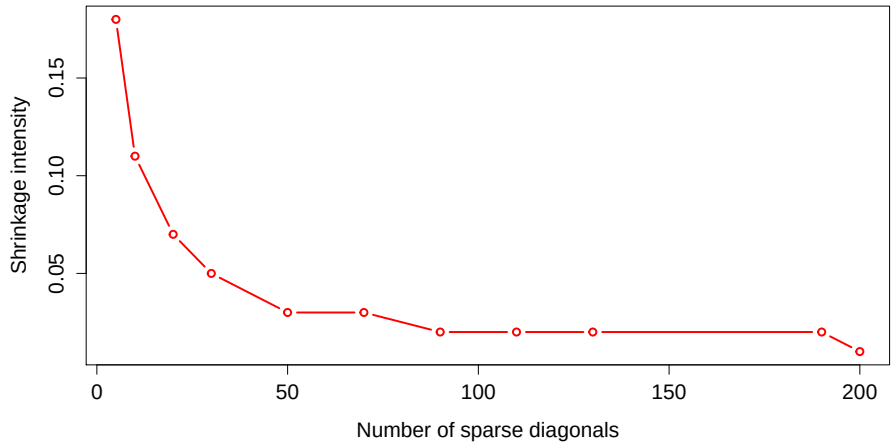

Figure 4.4: Effect of the number of sparse diagonals on optimal shrinkage intensity. True CNV size is set be 50 and total number of markers is 500.

As shown above, the number of diagonals d that we sparse the correlations also has a lot of effects on the estimation of correlation matrix. Denote bw as the bandwidth used for kernel aggregation method while d is the number of diagonals that create a sparse correlation matrix. We always don't know either where the CNV is or the true bandwidth of CNV. Basically we can choose any bandwidth for our kernel method. It is shown that the larger bandwidth we choose for kernel aggregation method would increase the power of detecting association between phenotype and intensities. After choosing a bandwidth, the number of diagnose d would include different correlation

information. Smaller d would get more sparse matrix while losing a lot more correlation information. Furthermore, d choosing also effects type one error. The banded sparse correlation matrix is more close to the whole sample correlation and includes more information when d gets larger. By Theorem 2, when the estimated correlation matrix gets much closer to the actual correlation matrix, the correlation approach would preserve appropriate type one error. Therefore, we aim to select optimal d to keep the correlation matrix sparse and also contain as much correlation information and meanwhile preserve type one error as possible.

Figure 4.5 illustrates the relationship between d and type one error for different bandwidth. For true CNV size of 50, we pick a set of bandwidths of 10,20,30,50,70,90 and 110, which include too small and too large bandwidth. Picking different number of sparse diagonals for each bandwidth, the optimal shrinkage intensity λs are chosen to be minimum to guarantee the positive definite structure of sparse correlation matrix. From the plot, the effect of number of sparse diagonals on type one error are all in the same trend for different bandwidths. Type one error goes down as increasing d until the type one error gets stable. Hence, d needs to be larger than some choice to preserve type one error. From simulation results, it seems to be safe to choose $d \geq 3bw$.

4.5 Simulation Results

I conducted simulation experiments in this section to evaluate the performance gain that the correlation method provides and the cost in terms of its ability to detect CNV-phenotype association. I first present the preservation of type one error of correlation approach in different settings. Also, I compare the correlation method

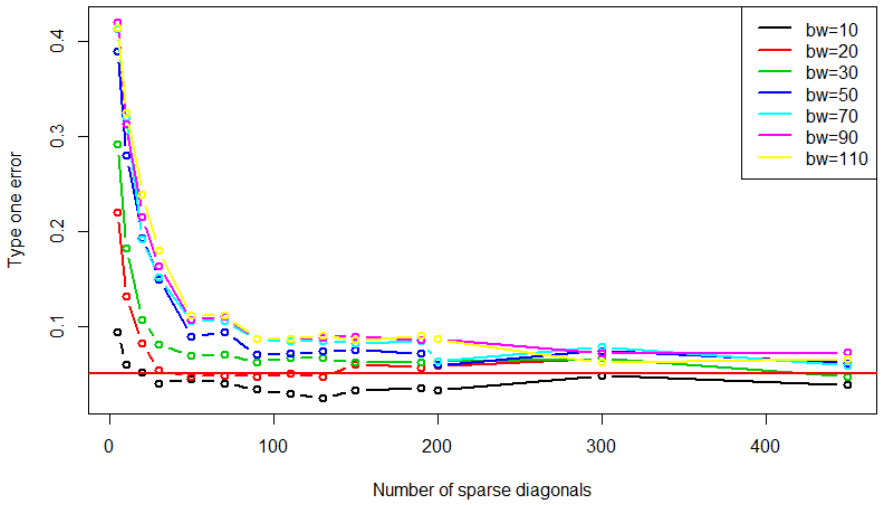

Figure 4.5: Effect of the number of sparse diagonals on preservation of type one error when choosing different bandwidth for kernel aggregation method on normal signed transformation. True CNV size is set be 50. The horizontal red line is for $\alpha = 0.05$

in different settings to permutation method in the distribution of the the estimated null distribution of the supremum statistic $T = \max_j\{T_j\}$ under no CNV-phenotype association. Then, comparison of the computing times applying different approach to conduct the kernel-based aggregation of marker-level association test will give us a measure of the performance gain. For all these simulations, I follow the spike-in data design fully described in Section 3.4. This procedure provides the simulation data similar to the real data. For the simulations presented here, I use a sample size of $n = 1,000$ and an total markers $J = 200$ for "large n, small J" setting; use a sample size of $n = 50$ and $J = 500$ assuming "small n, large J"; and $n = 300$ and $J = 500$ in the case of "large n, large J". Set signal-to-noise ratio of 2 for simulations.

4.5.1 Preservation of type one error

First of all, we are interested in the preservation of type one error of correlation approach in each setting described above. Preservation of type one error is one

important aspect to determine the performance of the method. In different settings, we all consider a genomic region in which individuals may have a CNV. The purpose of the analysis is to detect and locate such CNVs associated with a particular phenotype. The null hypothesis for our association test may hold in one of two ways: (1, "No CNV") no individuals with CNVs are present in the sample, or (2, "No association") individuals with CNVs are present in the sample, but the CNV does not affect the disease and thus dose not change the probability of developing the phenotype. The effect of transformations of p-values and association direction would be similar. In this section, I focus on the analysis for different transformations of p-values in signed case. Simulation results are shown in tables and figures.

4.5.1.1 "large n, small J" setting

For "large n, small J" setting, I set $n = 1000$ and $J = 200$, which is exactly the same setting as permutation approach. The results for the FWER analysis are summarized in Table 4.1 for different transformations of p-values in signed direction of associations.

Table 4.1 perfectly demonstrates that correlation method under such setting is guaranteed to preserves the correct type I error under both null hypothesis for different transformations in signed directions of association under linear model assumption between phenotype and genotype. This phenomenon is also illustrated graphically in Figure 4.6. Obviously, the correlation method to estimate the whole sample correlation is nice to apply since it preserves accurate type one error and p-values under the null appear to be uniformly distributed.

4.5.1.2 "small n, large J" setting

For "small n, large J" setting, I set $n = 50$ and $J = 500$ for simulation. As details described before, the rank of the estimated sample correlation does not depend on the

Table 4.1: Preservation of Type I error for correlation method in "large n, small J" setting for different transformations in signed direction of association with nominal $\alpha = .05$ in two possible settings for which the null hypothesis holds. The simulated genomic region contained 200 markers, 30 of which were spanned by a CNV with signal-to-noise ratio of 2. The CNV was present in either 0% or 50% of the 1000 samples, depending on the null hypothesis setting. A detailed description of the simulation data is given in Section 3.4.

	Signed None	Signed Normal	Signed Log
No CNV	0.041	0.042	0.046
No Association	0.043	0.046	0.046

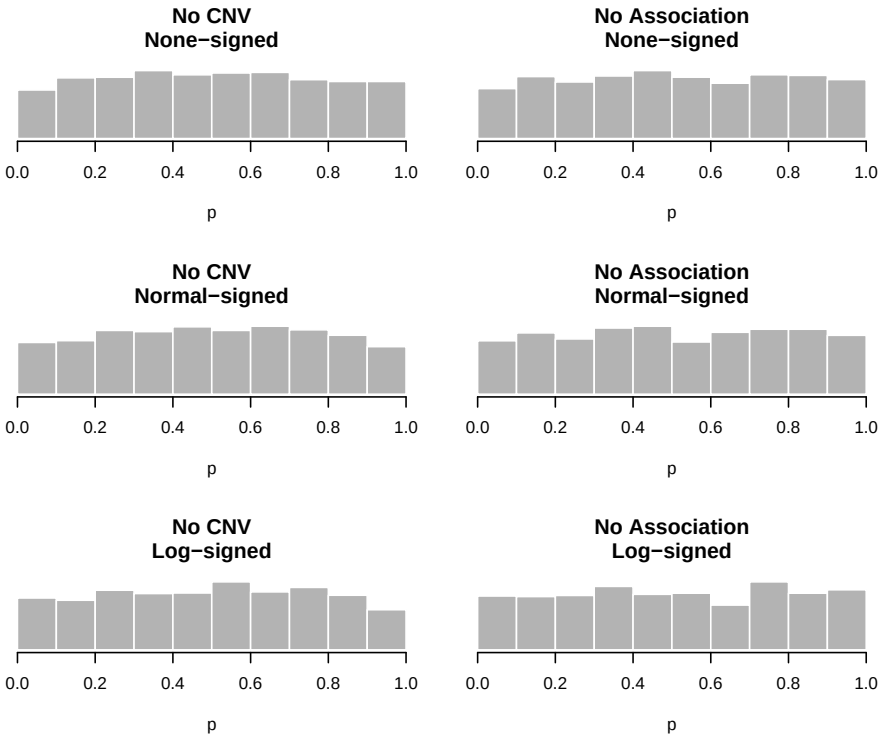

Figure 4.6: Ability of correlation approach under different transformations in signed directions of association for "Large n, small J" setting to maintain family-wise error rate under the two null scenarios.

number of markers but on the sample size n. Applying singular value decomposition, we simplify the correlation matrix and speed up the calculation. The results for the FWER analysis are concluded in Table 4.2 for signed direction of associations with

different transformations of p-values. And the distributions of p-values under two kinds of H_0 are shown as well in Figure 4.7.

Table 4.2: Preservation of Type I error for correlation method in "small n, large J" setting for different transformations and directions of association with nominal $\alpha = .05$ in two possible settings for which the null hypothesis holds. The simulated genomic region contained 500 markers, 30 of which were spanned by a CNV with signal-to-noise ratio of 2. The CNV was present in either 0% or 50% of the 50 samples, depending on the null hypothesis setting. A detailed description of the simulation data is given in Section 3.4.

	Signed None	Signed Normal	Signed Log
No CNV	0.049	0.049	0.048
No Association	0.046	0.047	0.048

From the table and figure, it is pretty good results in terms of the preservation of type I error and p-values under two null are close to uniform distribution. Therefore, although it is quite difficulty to estimate the whole sample correlation matrix using tradition way under "small n, large J" setting, SVD approach provides a perfect alternative to estimate the correlation matrix and conduct the kernel-base aggregation of marker-level association test.

4.5.1.3 "large n, large J" setting

Similarly, under "large n, large J" setting, I take $n = 300$ and the total number of markers $J = 500$. Applying shrinkage approach, it is necessary to select an optimal number of sparse diagonals and an appropriate shrinkage intensity. As illustrated in Section 4.4.3, the larger the number of sparse diagonal, the more information the sparse correlation matrix includes and thus the more accurate estimate of correlation

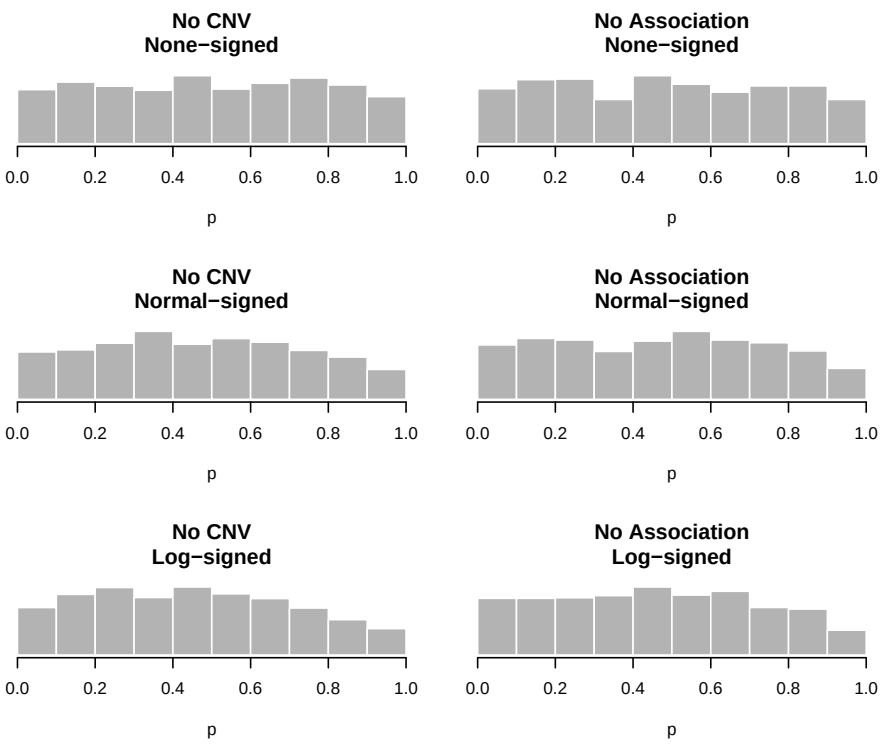

Figure 4.7: Ability of SVD approach under "small n, large J" setting for different transformations of p-values in signed association to maintain family-wise error rate under the two null scenarios.

matrix, which requires less shrinkage. It is suggested to pick $d \geq 3bw$ from simulations. Figure 4.5 shows that type one error tend to be stable when the number of sparse diagonals is larger than one specific number. Therefore, as we do not know the true CNV size, it will be better to pick a large d to preserve type I error and contain the important correlation part while choosing a smaller d to keep it sparse. For the simulation, I choose a bandwidth of 30 markers for kernel-based method and pick the number of sparse diagonal $d = 150$ and minimum shrinkage intensity $\lambda = 0.012$ to guarantee the sparse and positive-definite correlation matrix. The results of FWER analysis are present in Table 4.3 and Figure 4.8.

Under "large n, large J", a lot of approaches can not be applied and thus corre-

Table 4.3: Preservation of Type I error for applying shrinkage approach on correlation method for different transformations in signed association with nominal $\alpha = .05$ in two possible settings for which the null hypothesis holds. The simulated genomic region contained 500 markers, 30 of which were spanned by a CNV with signal-to-noise ratio of 2. The CNV was present in either 0% or 50% of the 300 samples, depending on the null hypothesis setting.

	Signed None	Signed Normal	Signed Log
No CNV	0.052	0.047	0.038
No Association	0.062	0.059	0.053

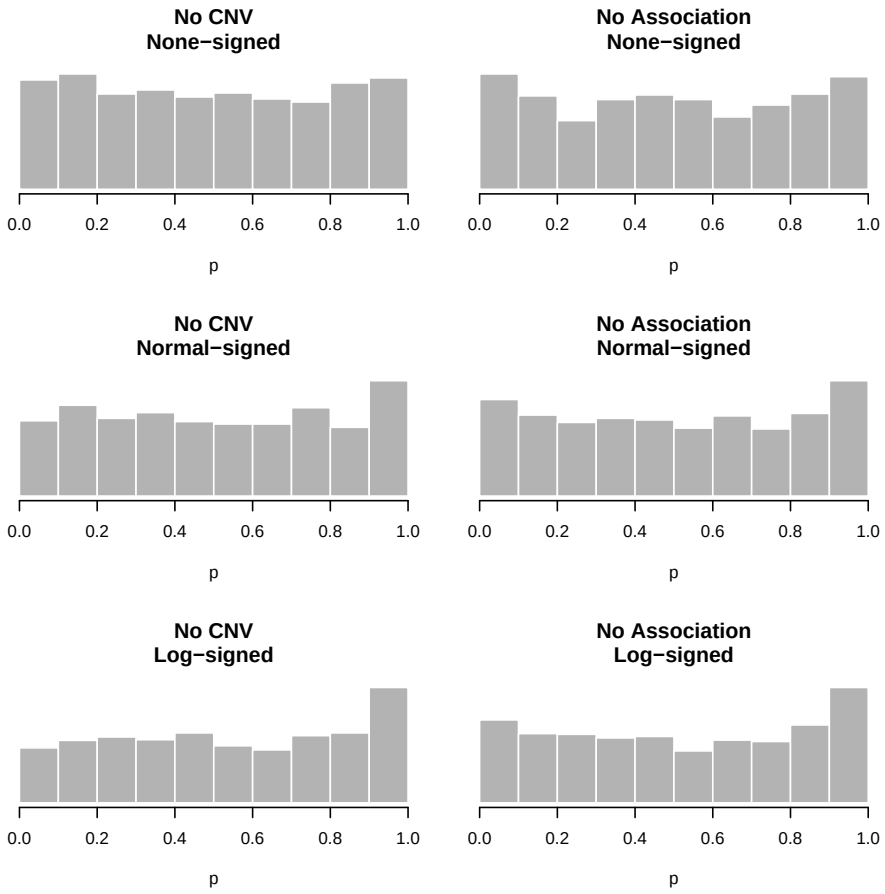

Figure 4.8: Ability of shrinkage approach under different transformations in signed association to maintain family-wise error rate under the two null scenarios.

lation matrix is complicated to estimate. In terms of preservation of type one error, the shrinkage method seems to be a nice choice to estimate the correlation matrix

under this setting when picking a proper d and shrinkage intensity λ.

4.5.2 Evaluating the estimated null distribution

It was clearly demonstrated that correlation approach in different settings successfully preserves type one error and thus they could be good to apply for estimation of correlation matrix and conduct kernel-base aggregation of association tests. Then we are interested in the estimation of the method under different settings. To see the estimate of the method, it would be appropriate to compare the estimated null distributions. I use the same settings as picked in last section for simulation. For each setting, I compare the null distributions for each setting when applying permutation method, the whole sample correlation estimate, SVD and shrinkage approach to the kernel-based aggregation of marker-level association test.

From Figure 4.9, Figure 4.10, Figure 4.11, the null distributions stay very close under different settings. Thus, the estimate and analysis results of kernel-based association test would be similar for all those methods and we are capable to use those approaches for each setting if not considering the performance. Besides, it is worth to note that the shape of the null distributions are still similar to each other even if we don't use the true CNV bandwidth.

4.5.3 Performance of correlation procedure

In the previous sections we showed that the estimated null distribution of correlation method works well as the permutation method and they all successfully preserve type one error. Thus, those correlation methods are good to use for any situation. Now, we

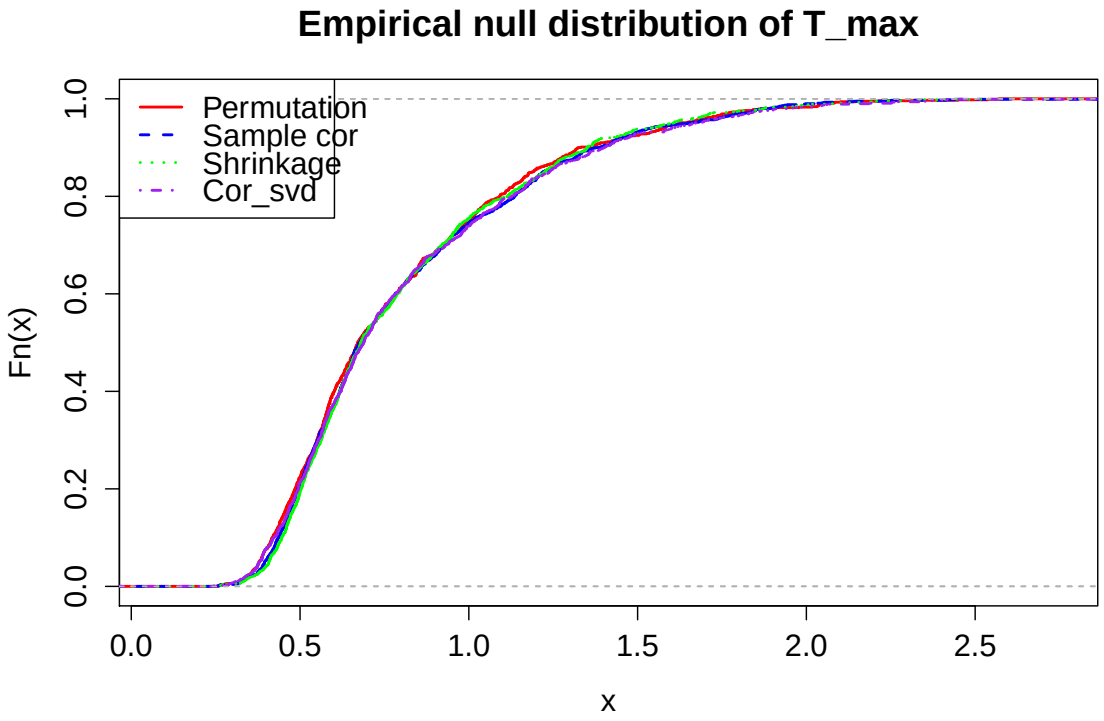

Figure 4.9: Comparison of null distribution through permutation and three correlation methods under "large n, small J" setting. The simulated genomic region contained 200 markers, 30 of which were spanned by a CNV with signal-to-noise ratio of 2. Population CNV frequency was presented in 50% of 1000 samples. Signed, normal transformation with bandwidth $bw = 20$ was used for kernel-based method.

are more interested in evaluating the operating characteristics of correlation methods. Unlike permutation method, the proposed correlation procedure involves the simulation of multivariate normal variables rather than the genotype or phenotype data and does not require repeated analysis of simulated data sets. The estimate of correlation matrix involving the observed data is calculated only once, and the evaluation of the null distribution given the estimated correlation matrix is trivial. Thus, the proposed correlation approach provides a much more efficient procedure to generate the null distribution than permutation approach. In this section we wish to assess the reduction in computing time achieved by this procedure at no cost in terms of false positive. We compare the performance of the correlation method to permutation

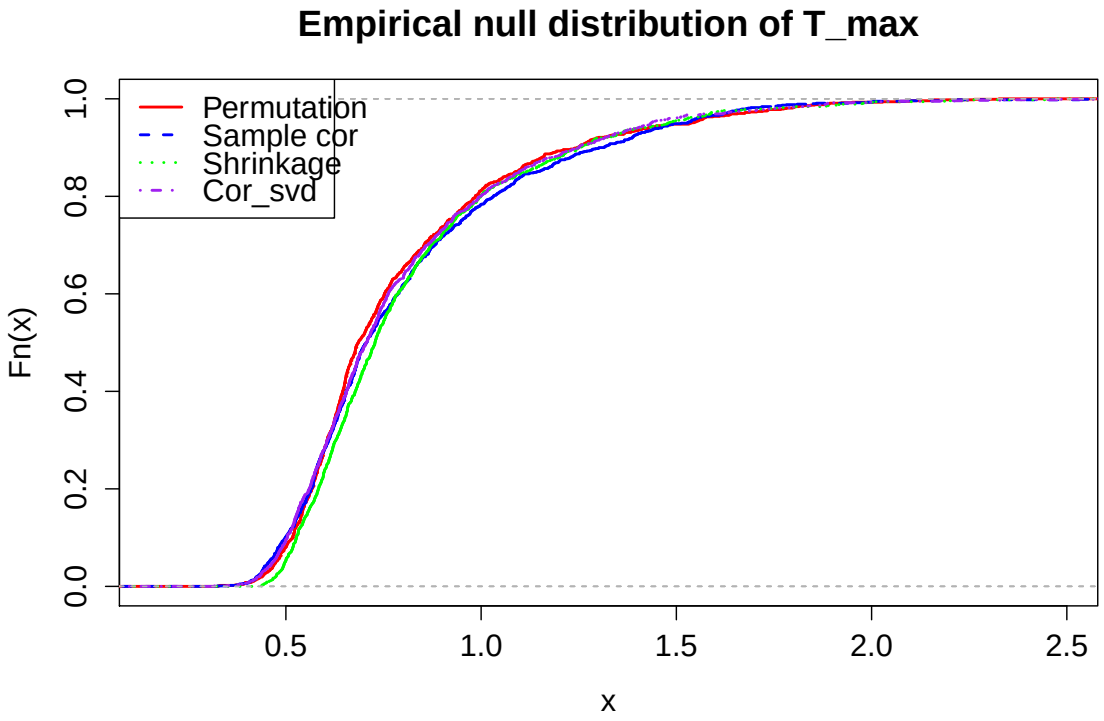

Figure 4.10: Comparison of null distribution through permutation and three correlation methods under "small n, large J" setting. The simulated genomic region contained 500 markers, 30 of which were spanned by a CNV with signal-to-noise ratio of 2. Population CNV frequency was presented in 50% of 50 samples. Signed, normal transformation with bandwidth $bw = 20$ was used for kernel-based method.

approach in kernel-based aggregation of marker-level association test. We would like to show the comparisons in terms of computation burdens and efficiency in low and high dimensions.

4.5.3.1 "large J" setting

As illustrated, correlation-based approach has extended to the high dimensions in efficient way. For "large J" setting, SVD approach is perfect for small sample size and shrinkage is nice for large sample size. We set the total number of markers to be 2000 for our simulation. When changing the sample size, the computation time

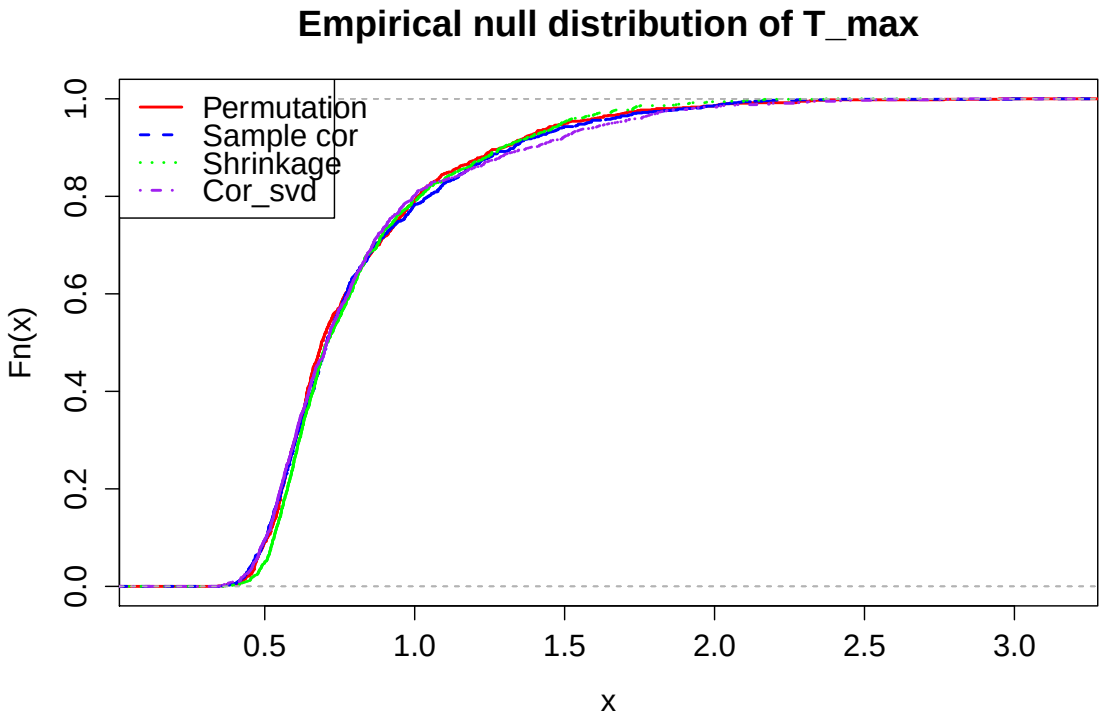

Figure 4.11: Comparison of null distribution through permutation and three correlation methods under "large n, large J" setting. The simulated genomic region contained 500 markers, 30 of which were spanned by a CNV with signal-to-noise ratio of 2. Population CNV frequency was presented in 50% of 300 samples. Signed, normal transformation with bandwidth $bw = 20$ was used for kernel-based method.

comparisons are presented in Figure 4.12. First of all, the black line stays a lot higher than the other three lines. Obviously, correlation approach performs much better than the permutation approach in terms of the computation burden. Furthermore, when the sample size gets larger, the gap between the black line ant the other three lines becomes wider. It makes sense since the sample size effects a lot on the computation time of permutation approach. Permutation approach would consume a lot more time as the sample size gets larger. More clear comparisons of three correlation methods under the same setting are shown in Figure 4.13. First, correlation method when estimating the whole correlation matrix runs longer time than SVD and shrinkage approach. It makes sense since the tradition correlation method estimates the whole

correlation matrix while SVD and shrinkage just provide estimate of part of the correlation matrix. In terms of the computation time, SVD approach is much better than shrinkage approach when the simple size is much smaller and meanwhile the shrinkage approach is getting better when the sample size gets larger.

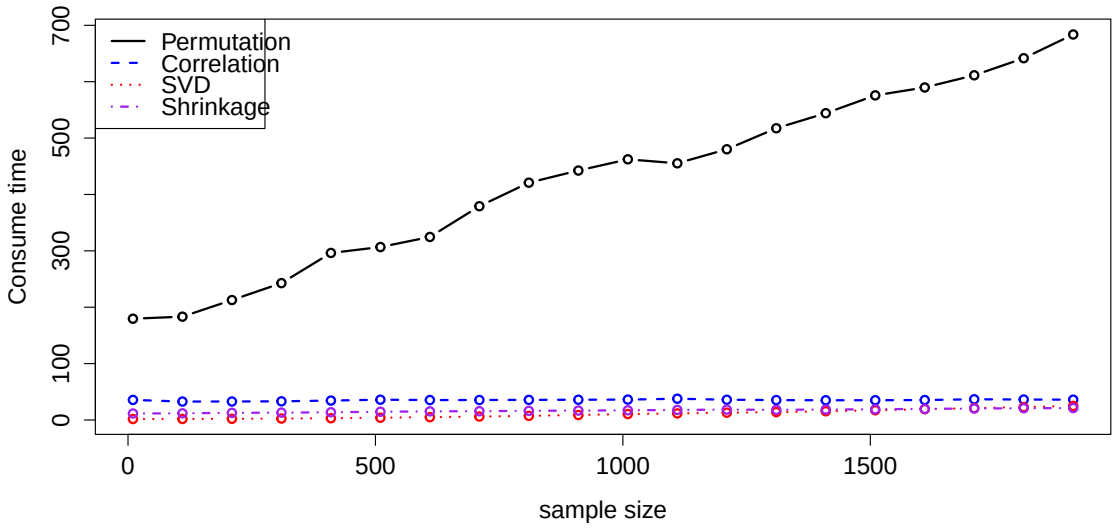

Figure 4.12: Comparison of computation time versus sample size between correlation methods and permutation approach. We set total number of markers to be 2000 and the total number of samples changes from 10 to 1910 by 100. CNV size are 30 markers.

4.5.3.2 "small J" setting

In "small J" setting, we set the total number of markers to be 200. While changing the sample size from 10 to 1910 by 100, the consuming time versus sample size is present in Figure 4.14. We see that the black line is always higher than the blue and the gap becomes larger and larger as the sample size gets larger. Thus, correlation approach is also much better than permutation approach even under low dimension. Besides, the consuming time for correlation method is more stable by sample size because the estimate of correlation matrix depends on the number of markers not the sample size. It is remarkable that he correlation method runs no more than one second here for

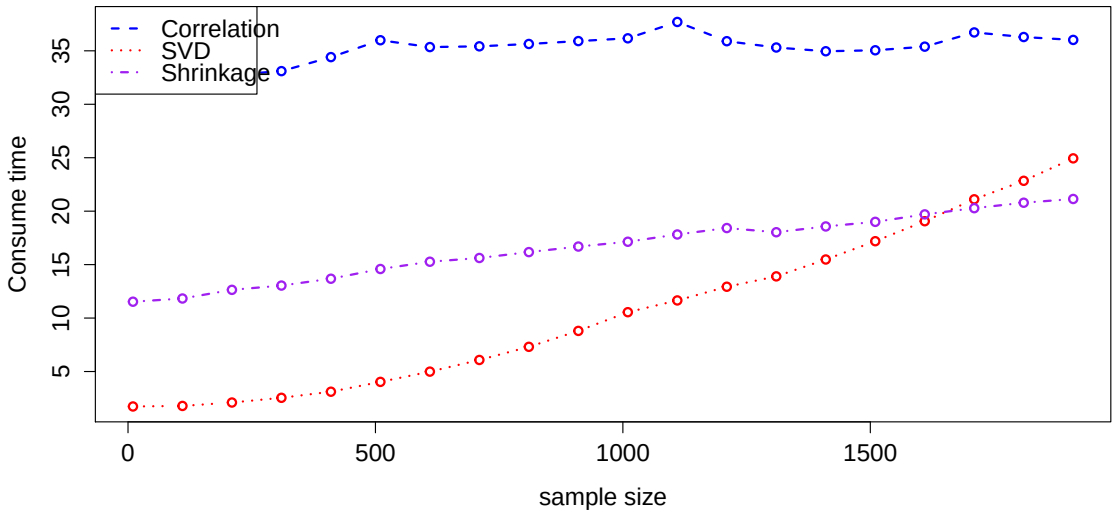

Figure 4.13: Comparison of computation time versus sample size among three kinds of correlation approaches. We set total number of markers to be 2000 and the total number of samples changes from 10 to 1910 by 100. CNV size are 30 markers.

situation of 200 total number of markers. Note that here the correlation method is giving the estimate of the whole correlation matrix.

This section demonstrates the performance of those methods and we would get the conclusion that correlation approach is a better choice than the permutation approach in terms of the running time in both low and high dimensions. Moreover, the SVD and shrinkage method are well applied for high dimension and consume less time than the tradition correlation method that gives the estimate of whole correlation matrix.

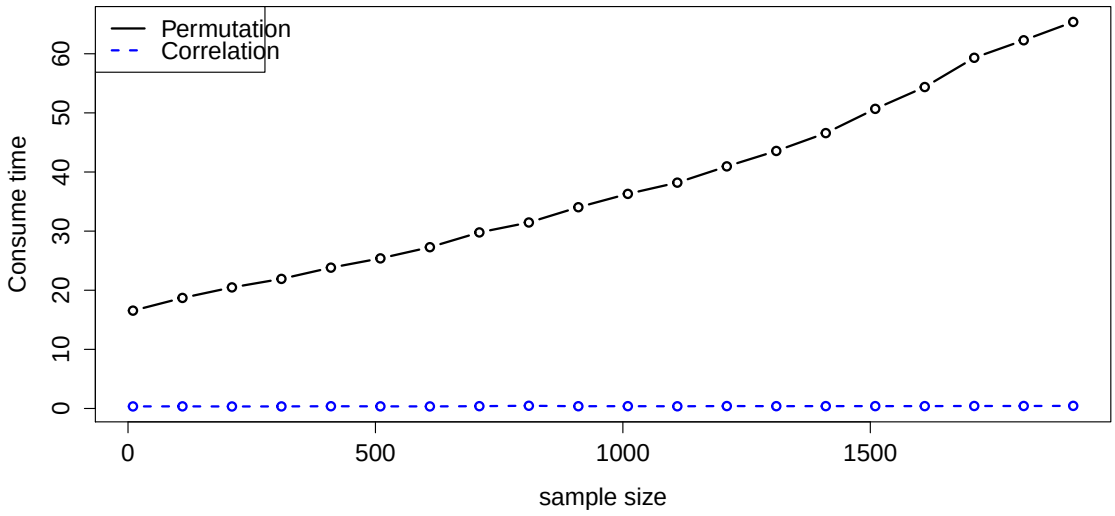

Figure 4.14: Comparison of computation time versus sample size between permutation approach and correlation method for low dimension. We set total number of markers to be 200 and the total number of samples changes from 10 to 1910 by 100. CNV size are 30 markers.

4.6 Gemcitabine study

In this section we apply the correlation approach to the real data and then compare it to permutation approach. We begin by describing the design of a pharmacogenomic study of gemcitabine, a commonly used treatment for pancreatic cancer. It is the same data that is analyzed for application of permutation approach. Then we analyze data applying the proposed correlation method on kernel-based aggregation association test.

The gemcitabine study was carried out on the Human Variation Panel, a model system consisting of cell lines derived from Caucasian, African-American and Han Chinese-American subjects (Coriell Institute, Camden, NJ). Gemcitabine cytotoxicity assays were performed at eight drug dosages (1000, 100, 10, 1, 0.1, 0.01, 0.001, and 0.0001 uM) [60]. Estimation of the phenotype IC50 (the effective dose that kills

50% of the cells) was then completed using a four parameter logistic model [61]. Marker intensity data for the cell lines was collected using the Illumina HumanHap 550K and HumanHap510S at the Genotyping Shared Resources at the Mayo Clinic in Rochester, MN, which consists of a total of 1,055,048 markers [62,63]. Raw data were normalized according to the procedure outlined in [64]. 172 cell lines (60 Caucasian, 53 African-American, 59 Han Chinese-American) had both gemcitabine cytotoxicity measurements and genome-wide marker intensity data. To illustrate the application of the kernel-based aggregation approach, we selected one chromosome (chromosome 3) from the genome-wide data. To control for the possibility of population stratification, which can lead to spurious associations, we used the method developed by [65], which uses a principal components analysis (PCA) to adjust for stratification. At each marker, a linear regression model was fit with PCA-adjusted IC50 as the outcome and intensity at that marker as the explanatory variable; these models produce the marker-level tests.

We analyzed these data using the kernel-based approach described in Section 2.2 with a bandwidth of 50 markers and the log transformation. Instead of permutation method on kernel, we demonstrate the correlation approach. As known, correlation method depends on the estimation of correlation matrix and there are a couple of choice for the different settings. For the real data involving with 70,542 markers for 172 cell lines, it is definitely a huge correlation matrix. It would be out of memory to calculate the complete sample correlation matrix. But the small sample size indicates that SVD will be a nice choice to estimate the correlation matrix. Also, we could try the shrinkage method with appropriate shrinkage intensity and the number of sparse diagonals. We pick $d = 300$ with minimum $\lambda = 0.03$ here. The results are shown in Figure 4.15. Note the presence of a peak at 102.6 Mb. The horizontal lines indicate the FWER-controlled, chromosome-wide significance threshold at the

$\alpha = 0.1$ level. These two cutoffs are close. It makes sense since the null distributions are close to each other shown in Section 4.5.2. As the figure indicates, there are both insufficient evidence in this study to establish a CNV association involving response to gemcitabine ($p = 0.158$ for SVD and $p = 0.204$ for shrinkage) after controlling the chromosome-wide FWER. Other choices of bandwidth and transformation would produce qualitatively similar, although somewhat less significant, results. Compared to the result of permutation in Figure 3.2 on the same real data with the same setting, these cutoffs are close to the cutoff of permutation approach, which is around 2.5. Note that the cutoff of shrinkage method is a little bit higher than the SVD and permutation approaches because shrinkage method losses correlation information to some extent. Therefore, correlation method and permutation approach provide similar results in terms of the performance. In terms of consuming time, SVD approach and shrinkage method run much faster than permutation, which requires resampling and runs a few hours. For this study, SVD just needs a couple of minutes to run the result and shrinkage require approximately 15 mins for kernel-based study plus extra time to get sparse correlation matrix and find out the optimal shrinkage intensity for positive-definite correlation. Therefore, SVD approach would the best choice for such data with much less sample size than the number of markers.

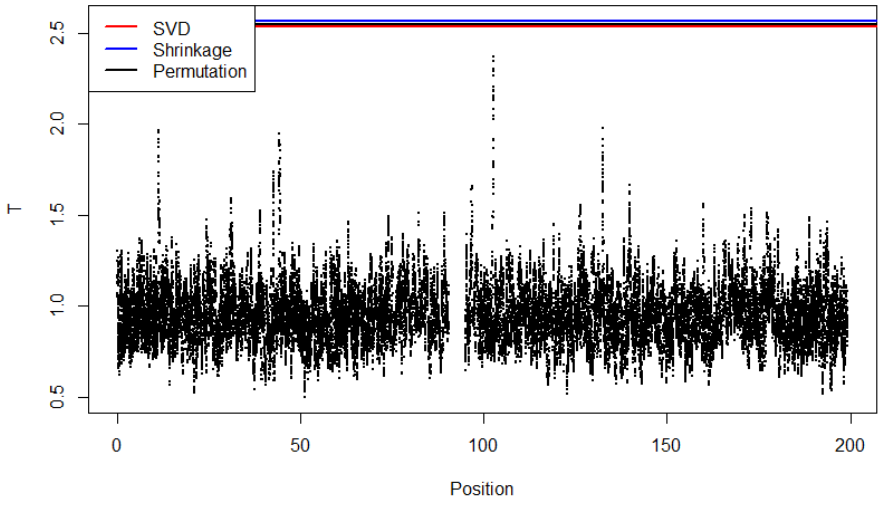

Figure 4.15: Analysis of the gemcitabine data (Chromosome 3) using the proposed correlation method. The kernel aggregations T_j are plotted against chromosomal position. The red line indicates the cutoff of SVD approach and the blue line shows the cutoff of shrinkage approach for chromosome-wide FWER significance at the $\alpha = .1$ level.

Chapter 5 Summary and Discussion

This dissertation is devoted to the analysis of CNV-phenotype association testing. Specially, we have reviewed the traditional and popular variant-level testing, which do "CNV calling" first for each individual and then carry out association test of whether individual with a CNV differ from individual without a CNV with respect to some phenotype. We focus on the marker-level testing, where we do the association test for every single marker first then determine CNV-phenotype associated regions by pooling test results across neighboring markers. Here in the dissertation we develop a kernel-based aggregation method for marker-level association test. Conducting such an association test, I propose a permutation approach and correlation method to estimate the null distribution of the test statistic.

In summary, permutation tests provide a robust and powerful method of testing statistical hypotheses that is intuitive and easy in practice. More importantly, it provides an accurate FWER control and does not rely on any model assumptions. However, the permutation approach has its own limitations. First, this approach is valid and widely used only under very mild conditions— complete exchangeability under null hypothesis as described above. Thus it may not be applicable when there are covariates or nuisance parameters [57]. Especially, the permutation distribution may not be appropriate when the analysis involves covariates that are correlated with both genotype and phenotype. Moreover, it becomes computationally demanding since the analysis needs to be repeated for each permuted dataset while creating the null distribution of test statistic. The kernel aggregation itself is very fast, but the need to carry out $\approx 1,000$ permutation tests for each marker may be highly computationally intensive, depending on the complexity of the marker-level test. Thus,

the computing problem is critical for permutation approach.

The proposed correlation method is demonstrated to offer large improvements in speed. First it does not involve repeated analyses of simulated datasets and is thus provides a substantial gain in speed with only a negligible loss in accuracy. Second, it does not require complete exchangeability and is thus widely applicable. Thus, correlation method appears to be a better choice than permutation in terms of the computation time. Moreover, extending this method to high dimension, we develop SVD and shrinkage method, which are presented to also preserve type one error and require much less running time.

The simulation studies of Section 3.4 address a limited-scale version of a larger question: how do marker-level test aggregation and variant-level testing compare for chromosome-wide and genome-wide analysis? This is an important question and deserves further study. In general, multiplicity is a thorny issue for CNV analyses, as the true location of CNVs are unknown and can overlap in a number of complicated ways. The issue of how many tests to carry out and adjust for is a challenging question for variant-level testing and a considerable practical difficulty in analysis. In contrast, aggregation of marker-level results avoids this issue altogether. We have shown that the proposed approach is both powerful at detecting CNV associations and rigorously controls the FWER at a genome-wide level — two rather appealing properties. However, future work applying the proposed method to larger, more complex settings is necessary.

In this dissertation, I have focused on continuous phenotype, with association testing performed using linear regression. The kernel-based aggregation method itself,

however, requires only p-values and can be extended to a more complicated marker-level tests assuming nonlinear, mixed-effects, or mixture models between intensity and phenotype. Furthermore, our simulations involve a very simple genetic scenario: a small segment of DNA in which a single CNV is either present or absent. It is important and valuable to understand the properties of marker-level approach in these simple cases. However, future research involving more complicated scenarios is also needed.

Appendices

A.Proof of Theorem 1

Proof of Theorem 1. Let $\mathcal{P}$ denote the set of all possible permutations of $\{y_i\}$, F_0 the CDF of $T_{\max}$ over $\mathcal{P}$, and F_0^{-1} its generalized inverse. Also, let $\phi(\mathbf{X}, \mathbf{y}) = 1$ if $T_{\max}(\mathbf{X}, \mathbf{y}) > F_0^{-1}(1-\alpha)$ and 0 otherwise.

Now, note that under the null hypothesis that $\mathbf{x}_i$ and y_i are independent,

$$\begin{aligned} P(\mathbf{X}, \mathbf{y}) &= \prod_i P(\mathbf{x}_i, y_i) \\ &= \prod_i P(\mathbf{x}_i) P(y_i) \\ &= P(\mathbf{X}, \mathbf{y}^*) \end{aligned}$$

for all $y^* \in \mathcal{P}$. Thus, $\mathbb{E}_0 \phi(\mathbf{X}, \mathbf{y}^*)$ is a constant for all $\mathbf{y}^*$ and

$$\begin{aligned} \mathbb{E}_0 \left\{ \phi(\mathbf{X}, \mathbf{y}) \right\} &= \frac{1}{n!} \sum_{\mathbf{y}^* \in \mathcal{P}} \mathbb{E}_0 \phi(\mathbf{X}, \mathbf{y}^*) \\ &= \mathbb{E}_0 \frac{1}{n!} \sum_{\mathbf{y}^* \in \mathcal{P}} \phi(\mathbf{X}, \mathbf{y}^*) \\ &\leq \alpha, \end{aligned}$$

where the term inside the expectation in the second line is less than or equal to α for all $\mathbf{X}$ and $\mathbf{y}$ by the construction of the test. □

B. Proof of Theorem 2

Proof of Theorem 2. Assume simple linear regression $y = \beta_0 + \beta_{1j}x_j + \epsilon$ between intensities x_j and y for subjects is, where ϵ has a normal distribution with mean 0 and standard deviation σ. By standardization, we define $x_{ij} - \bar{x_j} = x'_{ij}sd_{x_{ij}}$ and $y_i - \bar{y} = y'_i sd_{y_i}$, where we denote standardized x'_{ij} and y'_i. Under the simple linear regression model, we have the z-statistics under H_0:

$$\begin{aligned} Z_j &= \frac{\hat{\beta_{1j}} - 0}{s.e.} \\ &= \frac{\frac{\sum(x_{ij}-\bar{x_j})(y_i-\bar{y})}{\sum(x_{ij}-\bar{x_j})^2}}{\hat{\sigma}\sqrt{1/\sum(x_{ij}-\bar{x_j})^2}} \\ &= \frac{\sum sd_{x_{ij}}x'_{ij}sd_{y_i}y'_i}{\hat{\sigma}(n-1)sd^2_{x_{ij}}}\sqrt{(n-1)sd^2_{x_{ij}}} \\ &= \frac{\sum x'_{ij}y'_i}{\sqrt{(n-1)}}, \end{aligned}$$

where $\sum(x_{ij} - \bar{x_j})^2 = (n-1)sd^2_{x_{ij}}$ and also $\hat{\sigma} = sd_{y_i}$.

Now, correlation between two p-values is equivalent to the correlation between the two corresponding z statistics. Therefore, correlation between z statistics of two markers is simplified by (4.2),

$$\begin{aligned} cor(z_j, z_k) &= cor(\frac{\sum x'_{ij}y'_i}{\sqrt{(n-1)}}, \frac{\sum x'_{ik}y'_i}{\sqrt{(n-1)}}) \\ &= cor(\sum x'_{ij}y'_i, \sum x'_{ik}y'_i) \\ &= cor(\sum x'_{ij}, \sum x'_{ik}), \end{aligned}$$

since under the null hypothesis that x_{ij} and y_i are independent.

By (4.2) again,

$$\begin{aligned} cor(z_j, z_k) &= cor(\sum x'_{ij}, \sum x'_{ik}) \\ &= cor(\mathbf{x}_j, \mathbf{x}_k) \end{aligned}$$

Thus,correlation structure among p-values under H_0 exactly equals the correlation matrix of the intensities among markers. □

C. R code for Permutation Method

```
####kernel-based association test


## Input: X (matrix of intensity values)
##        y (vector of phenotypes)
##        FUN (If P and S are not supplied, must supply FUN
##             a function which carries out the marker-level testing;
##             must return vector of p-values or a list with components
##             'p' and 's' if signed aggregation is to be carried out)
##        bw (bandwidth)
##        pos (position of markers on chromosome)
##        trans (transformation)
##        test (if TRUE, calculates F0)
##        N (Number of permutations)
##        ... (arguments passed to FUN)
## Output: T (test statistic)
##         p (p-value of t)
##         Tmax
kbag <- function(obj, ...) UseMethod("kbag")
kbag.permTest <- function(obj, ...)
{
  if (attr(obj, "signed")) kbag.numeric(p=obj$p, s=obj$s, P=obj$P, S=obj$S, ...)
  else kbag.numeric(p=obj$p, P=obj$P, ...)
}
kbag.function <- function(FUN, X, y, test=TRUE, N=1000, showProgressBar=TRUE, ...)
{
  ## Evaluate FUN
```

```
FUN.args <- as.list(formals(FUN))
dots <- list(...)
matched <- names(dots)[names(dots) %in% names(FUN.args)]
if (length(matched)) FUN.args[matched] <- dots[matched]
FUN.args[[1]] <- X
FUN.args[[2]] <- y
fun.val <- do.call(FUN, FUN.args)
signed <- FALSE
s <- S <- P <- NULL
if (is.numeric(fun.val)) {
  p <- fun.val
} else if (is.list(fun.val)) {
  p <- fun.val$p
  if ("s" %in% names(fun.val)) {
    s <- fun.val$s
    signed <- TRUE
  }
} else stop("FUN returns unrecognized format")

## Calculate P, S
if (test) {
  P <- matrix(NA, nrow=N, ncol=ncol(X))
  if (signed) S <- P
  for (i in 1:N) {
    FUN.args[[2]] <- sample(y)
    res <- do.call(FUN, FUN.args)
    if (!signed) P[i,] <- res else {
```

```
        P[i,] <- res$p
        S[i,] <- res$s
      }
      if (showProgressBar) displayProgressBar(i,N)
    }
  }

  if (signed) kbag.numeric(p, s, P=P, S=S, test=test, ...) else kbag.numeric(p, P=
}

kbag.numeric <- function(p, s, bw, X, N=1000, P, S, pos=1:length(p), trans=c("log"
{
  ## Aggregate
  signed <- if (missing(s)) FALSE else TRUE
  if (signed & missing(trans)) trans <- "normal"
  trans <- match.arg(trans)
  if (missing(bw)) stop("You must supply a bandwidth")

  if (signed) {
    if (trans=="none") x <- s*(1-p)
    if (trans=="normal") x <- qnorm((1+s*(1-p))/2)
    if (trans=="log") x <- -s*log(p)
  } else {
    if (trans=="none") x <- 1-p
    if (trans=="normal") x <- qnorm(1-p)
    if (trans=="log") x <- -log(p)
  }
```

```
out <- .C("KBAGN", double(length(p)-2*(bw-1)), integer(length(p)-2*(bw-1)), inte
T <- out[[1]]
names(T) <- pos[out[[2]]]
Tmax <- if (signed) max(abs(T)) else max(T)

## Calculate F0, test
#########This part will be different when applying different approaches to get F
if (test) {
  if (missing(P)) {
    if (missing(X)) stop("If test=TRUE, must supply either X (matrix of intensit
    Sigma <- cor(X)
     require(mvtnorm)
    Z <- rmvnorm(N, sigma=Sigma, method="svd")
    if (signed) {
      P <- 2*pnorm(-abs(Z))
      S <- sign(Z)
    } else P <- 1-pchisq(Z^2,1)

  }
  F0 <- if (signed) getF0(P, S, bw=bw, pos=pos, trans=trans) else getF0(P, bw=bw
  p.value <- 1-F0(Tmax)
} else {
  p.value <- NULL
  F0 <- NULL
}
```

```
  ## Return
  structure(list(Tmax=Tmax, p=p.value, F0=F0, T=T, signed=signed, bw=bw, trans=tra
}

## Obtains null distribution for Tmax using permutation testing
getF0 <- function(P, S, bw, pos, trans)
{
  N <- nrow(P)
  x <- numeric(N)
  if (missing(S)) for (i in 1:N) x[i] <- kbag(P[i,], bw=bw, pos=pos, trans=trans,
  else for (i in 1:N) x[i] <- kbag(P[i,], S[i,], bw=bw, pos=pos, trans=trans, test
  ecdf(x)
}



###Simulation data
## n = number of subjects
## g = frequency of CNV
## m = number of SNPs / CNV
## snr = signal-to-noise ratio
## J = # of markers
## pen = penetrance
## standardized = standardize X and y?
genData <- function(n, g, m, snr=0.8, delta=1, J=200, pen, noise.type=c("spike","m
  {
    noise.type <- match.arg(noise.type)
```

```
if (m%%2!=0) stop("m must be even")

## Generate y, z
if (missing(pen))
  {
    z <- rbinom(n,1,g)
    y <- genY(n=n,delta=delta,z=z)
  }
else
  {
    if (n%%2!=0) stop("n must be even in a case-control study")
    y <- c(rep(0,n/2),rep(1,n/2))
    p <- numeric(n)
    p[y==1] <- g*pen[2]/(g*pen[2]+(1-g)*pen[1])
    p[y==0] <- g*(1-pen[2])/(g*(1-pen[2])+(1-g)*(1-pen[1]))
    z <- rbinom(n,1,p)
  }

## Generate X
X <- Z <- matrix(0,nrow=n,ncol=J)
j <- c(-(J/2):-1,1:(J/2))
Z[z==1,abs(j) <= m/2] <- snr

if (noise.type=="spike") {
  if (is.null(attr(R,"sd"))) {
    sd.r <- sd(as.numeric(R))
  } else sd.r <- attr(R,"sd")
```

```
      a <- sample(1:(ncol(R)-J),1)
      markers <- sample(1:ncol(R),J)
      noise <- as.numeric(R[sample(1:nrow(R),n,replace=TRUE),markers]/sd.r)
    } else {
      E <- matrix(rbinom(J*n,size=1,prob=.3),ncol=J)
      R1 <- matrix(rdex(J*n,0,1),ncol=J)
      R2 <- matrix(rnorm(J*n,0,1))
      noise <- E*R1+(1-E)*R2
    }
    X <- Z + noise

    val <- list(y=y,X=X,Z=Z,z=z)
    if (noise.type=="spike") {
      start <- sample(1:length(pos), 1)
      val$pos <- pos[start:(start+J-1)]
    }
    if (standardized) {
      val$X <- standardizeX(val$X)
      val$y <- standardizeY(val$y)
    }
    val
  }

genY <- function(n,delta,z,pen)
  {
    return(rnorm(n,mean=delta*z))
    ##yy <- rbinom(n,size=1,prob=pen[z+1])
```

```
    ##return(list(y=y,yy=yy))
  }


####standardize data
standardizeX <- function(X)
{
  n <- nrow(X)
  center <- colMeans(X)
  X.c <- sweep(X,2,center)
  scale <- sqrt(apply(X.c,2,crossprod)/(n-1))
  sweep(X.c,2,scale,"/")
}
standardizeY <- function(y){(y-mean(y))/sd(y)}


#####linear regression between intensity and phenotype


mlt <- function(XX, yy, type=c("continuous", "discrete"), signed=TRUE, return.line
{
  type <- match.arg(type)
  if (type=="continuous") {
    n <- length(yy)
    if (standardized) {
      b <- crossprod(XX,yy)/n
      R <- yy - sweep(XX,2,b,"*")
      t. <- b/(sqrt(apply(R,2,crossprod)/(n-2)/n))
    } else {
      meany <- mean(yy)
```

```
      y <- yy - meany
      meanx <- apply(XX,2,mean)
      X <- t(t(XX) - meanx)
      Xy <- crossprod(X,y)
      XX <- apply(X,2,crossprod)
      b <- as.numeric(Xy/XX)
      if (return.line) a <- meany-b*meanx
      R <- y - t(t(X)*b)
      t. <- b/(sqrt(apply(R,2,crossprod)/(n-2)/XX))
    }
  }
  if (type=="discrete") {
    n <- length(yy)
    fit <- lm(XX~yy)
    MSE <- apply(fit$residuals,2,crossprod)/(n-2)
    SXX <- crossprod(yy-mean(yy))
    SE <- sqrt(MSE/SXX)
    b <- fit$coef[2,]
    if (return.line) a <- fit$coef[1,]
    t. <- b/SE
  }
  p <- 2*pt(-abs(t.),n-2)
  if (!return.line & !signed) val <- p else val <- list(p=p)
  if (return.line) val <- append(val,list(b=b, a=a))
  if (signed) val <- append(val,list(s=sign(t.)))
  val
}
```

```
####plot kbag result####
plot.kbag <- function(x,F0,alpha=.05,pch=19,cex=.1,ylim,...)
  {
    if (!missing(F0)) x$F0 <- F0
    T <- x$T
    Position <- as.numeric(names(x$T))
    if (is.null(x$F0)) {
      if (missing(ylim)) ylim <- range(T)
      plot(Position,T,pch=pch,cex=cex,ylim=ylim,...)
    } else {
      cutoff <- quantile(x$F0,1-alpha)
      if (x$signed) {
        ylim <- range(c(T,-cutoff,cutoff))
        plot(Position,T,pch=pch,cex=cex,ylim=ylim,...)
        abline(h=cutoff,col="red",lwd=2)
        abline(h=-cutoff,col="red",lwd=2)
      } else {
        ylim <- range(c(T,cutoff))
        plot(Position,T,pch=pch,cex=cex,ylim=ylim,...)
        abline(h=cutoff,col="red",lwd=2)
      }
    }
  }

## Standardized
Data <- genData(n=300, g=0.5, m=50,delta=0, standardized=TRUE)
```

```
mlt(Data$X, Data$y)$p[1:10] ## Same as below
mlt(Data$X, Data$y, standardized=TRUE)$p[1:10] ## Same as above
fit  <- kbag(mlt, Data$X, Data$y, bw=30, standardized=TRUE)
plot(fit)



###Demo
###permutation-based###
Data <- genData(n=300, g=0.5, m=50,delta=0)
fit  <- kbag(mlt, Data$X, Data$y, bw=30, test=FALSE)
fit  <- kbag(mlt, Data$X, Data$y, bw=30)
plot(fit)

## P
Data <- genData(n=300, g=0.5, m=50, delta=0, standardized=TRUE)
P <- S <- matrix(NA, nrow=1000, ncol=200)
for (i in 1:1000) {
  res <- mlt(Data$X, sample(Data$y), standardized=TRUE)
  P[i,] <- res$p
  S[i,] <- res$s
  displayProgressBar(i, 1000)
}
res <- mlt(Data$X, Data$y, standardized=TRUE)
fit <- kbag(res$p, res$s, P=P, S=S, bw=30)
plot(fit)
fit <- kbag(res$p, res$s, P=P, S=S, bw=50, trans="none")
plot(fit)
```

D. R code for Correlation Method

```
#####Basic correlation method to get the whole sample correlation matrix#####


## Single example -- "manual"
Data <- genData(n=300, g=0.5, m=50, delta=0, snr=20)
res <- mlt(Data$X, Data$y)
Sigma <- cor(Data$X)
Z <- rmvnorm(1000, sigma=Sigma, method="svd")
P <- 2*pnorm(-abs(Z))
S <- sign(Z)
fit <- kbag(res$p, res$s, P=P, S=S, bw=30)
plot(fit)

## Single example -- "automatic"
Data <- genData(n=300, g=0.5, m=50, delta=0, snr=20)
res <- mlt(Data$X, Data$y)
fit <- kbag(res$p, res$s, bw=30, X=Data$X)
plot(fit)

## Simulation
N <- 500
p <- numeric(N)
for (i in 1:N) {
  Data <- genData(n=300, g=0.5, m=50, delta=0, snr=20)
  res <- mlt(Data$X, Data$y)
  fit <- kbag(res$p, res$s, bw=30, X=Data$X)
  p[i] <- fit$p
```

```
  displayProgressBar(i, N)
}
hist(p, col="gray", breaks=c(0,1,.01), border="white")



#####shrinkage approach
diags <- function(XX, d) {
  X <- standardizeX(XX)
  n <- nrow(X)
  p <- ncol(X)
  B <- matrix(0, p, d)
  for (i in 1:p) {
    for (j in 1:d) {
      if (i+j > p) break
      B[i,j] <- crossprod(X[,i], X[,(i+j)])/n
    }
  }
  B
}

smoothB1 <- function(B,alpha) {
  n <- nrow(B)
  p <- ncol(B)
  BB <- matrix(0, n, p)
  for (i in 1:n) {
    for (j in 1:p) {
```

```
      BB[i,j] <- B[i,j]*(1-alpha)^j
    }
  }
  BB
}

multiz3h <- function(n=1000,XX, d,alpha) {
  t<-ncol(XX)
  require(Matrix)
  require(psych)
  B <- diags(XX, d)
  BB<-smoothB1(B,alpha=alpha)
  S <- bandSparse(nrow(BB), k=0:ncol(BB), diag=cbind(rep(1, nrow(BB)),BB), symmetr
  cholS <- chol(S)
  z<- array(rnorm(t*n),c(n,t)) %*% cholS  ##see str(mvsamples)
  z<-as.matrix(z)
  return(z)     ####give n*t matrix
}



###find the best shrinkage intensity
aa<-seq(0.1,0.15,0.01)
l<-length(aa)
N=500
s<-array(data<-0,dim=c(l,N),dimnames=list(aa,1:N))
for (i in 1: l)
{
```

```
  for (j in 1:N)
  {
    Data <- genData(n=50,g=0.5,m=100,delta=0,snr=2,J=500)
    B <- diags(Data$X, 10)
    BB<-smoothB1(B,alpha=aa[i])  ###alpha=0.03 seems okay
    S <- bandSparse(nrow(BB), k=0:ncol(BB), diag=cbind(rep(1, nrow(BB)),BB), symme
    ss<-as.matrix(S)
    s[i,j]<-sum(eigen(ss)$values < 0)

  }
  displayProgressBar(i, l)
}
S<-apply(s,1,sum)

#####Getting F0 and only change one part of kbag function

 Z<-multiz3h(n=N,X,d,alpha)
 ###Or:
 Data <- genData(n=300,g=0,m=30,delta=0,snr=2,J=500)
  res <- mlt(Data$X, Data$y)
  fit <- kbag2(res$p, bw=30,d=150,alpha=0.012, X=Data$X, trans="none")



#####Applying SVD on kbag function
## Single example -- "manual"
Data <- genData(n=50, g=0.5, m=30, delta=0, snr=2,standardized=TRUE, J=500)
nn <- N<-1000
```

```
X<-Data$X
n=dim(X)[1]
X <- standardizeX(X)
ZZ <- matrix(rnorm((n-1)*nn), nn, n-1)
SVD <- svd(X, nu=0, nv=n-1)
A <- sweep(SVD$v, 2, SVD$d[1:(n-1)], "*")/sqrt(n)
Z <- tcrossprod(ZZ, A)
P <- 2*pnorm(-abs(Z))
S <- sign(Z)
fit <- kbag(res$p, res$s, P=P, S=S, bw=30)
##Or
fit <- kbag(res$p, res$s, bw=30, X=Data$X)
plot(fit)
```

Bibliography

[1] L. Kruglyak, D. Nickerson, *et al.*, "Variation is the spice of life," *Nature genetics*, vol. 27, no. 3, pp. 234–235, 2001.

[2] B. Maher, "The case of the missing heritability," *Nature*, vol. 456, no. 7218, pp. 18–21, 2008.

[3] A. Iafrate, L. Feuk, M. Rivera, M. Listewnik, P. Donahoe, Y. Qi, S. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature genetics*, vol. 36, no. 9, pp. 949–951, 2004.

[4] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi, *et al.*, "Large-scale copy number polymorphism in the human genome," *Science*, vol. 305, no. 5683, pp. 525–528, 2004.

[5] L. Feuk, A. Carson, and S. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[6] S. McCarroll, F. Kuruvilla, J. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. Shapero, P. de Bakker, J. Maller, A. Kirby, *et al.*, "Integrated detection and population-genetic analysis of snps and copy number variation," *Nature genetics*, vol. 40, no. 10, pp. 1166–1174, 2008.

[7] E. Tuzun, A. Sharp, J. Bailey, R. Kaul, V. Morrison, L. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, *et al.*, "Fine-scale structural variation of the human genome," *Nature genetics*, vol. 37, no. 7, pp. 727–732, 2005.

[8] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, *et al.*, "Global variation in copy number in the human genome," *nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[9] A. Sharp, D. Locke, S. McGrath, Z. Cheng, J. Bailey, R. Vallente, L. Pertz, R. Clark, S. Schwartz, R. Segraves, *et al.*, "Segmental duplications and copy-number variation in the human genome," *The American Journal of Human Genetics*, vol. 77, no. 1, pp. 78–88, 2005.

[10] J. Freeman, G. Perry, L. Feuk, R. Redon, S. McCarroll, D. Altshuler, H. Aburatani, K. Jones, C. Tyler-Smith, M. Hurles, *et al.*, "Copy number variation: new insights in genome diversity," *Genome research*, vol. 16, no. 8, pp. 949–961, 2006.

[11] S. Levy, G. Sutton, P. Ng, L. Feuk, A. Halpern, B. Walenz, N. Axelrod, J. Huang, E. Kirkness, G. Denisov, *et al.*, "The diploid genome sequence of an individual human," *PLoS biology*, vol. 5, no. 10, p. e254, 2007.

[12] C. Bridges, "The bar" gene" a duplication," *Science*, vol. 83, no. 2148, p. 210, 1936.

[13] H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. Gudnadottir, *et al.*, "A common inversion under selection in europeans," *Nature genetics*, vol. 37, no. 2, pp. 129–137, 2005.

[14] R. Redon, S. Ishikawa, K. Fitch, L. Feuk, G. Perry, T. Andrews, H. Fiegler, M. Shapero, A. Carson, W. Chen, *et al.*, "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[15] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, *et al.*, "Strong association of de novo copy number mutations with autism," *Science*, vol. 316, no. 5823, pp. 445–449, 2007.

[16] T. Walsh, J. McClellan, S. McCarthy, A. Addington, S. Pierce, G. Cooper, A. Nord, M. Kusenda, D. Malhotra, A. Bhandari, *et al.*, "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia," *Science's STKE*, vol. 320, no. 5875, p. 539, 2008.

[17] D. Zhang, L. Cheng, Y. Qian, N. Alliey-Rodriguez, J. Kelsoe, T. Greenwood, C. Nievergelt, T. Barrett, R. McKinney, N. Schork, *et al.*, "Singleton deletions throughout the genome increase risk of bipolar disorder," *Molecular psychiatry*, vol. 14, no. 4, pp. 376–380, 2008.

[18] J. Barber, C. Joyce, M. Collinson, J. Nicholson, L. Willatt, H. Dyson, M. Bateman, A. Green, J. Yates, and N. Dennis, "Duplication of 8p23. 1: a cytogenetic anomaly with no established clinical significance.," *Journal of medical genetics*, vol. 35, no. 6, pp. 491–496, 1998.

[19] J. Engelen, U. Moog, J. Evers, H. Dassen, J. Albrechts, and A. Hamers, "Duplication of chromosome region 8p23. 1âĘŠ p23. 3: a benign variant?," *American journal of medical genetics*, vol. 91, no. 1, pp. 18–21, 2000.

[20] B. Trask, C. Friedman, A. Martin-Gallardo, L. Rowen, C. Akinbami, J. Blankenship, C. Collins, D. Giorgi, S. Iadonato, F. Johnson, *et al.*, "Members of the olfactory receptor gene family are contained in large blocks of dna duplicated polymorphically near the ends of human chromosomes.," *Human molecular genetics*, vol. 7, no. 1, p. 13, 1998.

[21] B. Riley, M. Williamson, D. Collier, H. Wilkie, and A. Makoff, "A 3-mb map of a large segmental duplication overlapping the [alpha] 7-nicotinic acetylcholine receptor gene (chrna7) at human 15q13-q14," *Genomics*, vol. 79, no. 2, pp. 197–209, 2002.

[22] G. Ehret, P. Munroe, K. Rice, M. Bochud, A. Johnson, D. Chasman, A. Smith, M. Tobin, G. Verwoert, S. Hwang, *et al.*, "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, 2011.

[23] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer, and P. Lichter, "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances," *Genes, chromosomes and cancer*, vol. 20, no. 4, pp. 399–407, 1997.

[24] S. Zöllner and T. Teslovich, "Using gwas data to identify copy number variants contributing to common complex diseases," *Statistical Science*, vol. 24, no. 4, pp. 530–546, 2009.

[25] N. P. Carter, "Methods and strategies for analyzing copy number variation using dna microarrays," *Nature genetics*, vol. 39, pp. S16–S21, 2007.

[26] D. Peiffer, J. Le, F. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. Shaw, J. Belmont, *et al.*, "High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping," *Genome research*, vol. 16, no. 9, pp. 1136–1148, 2006.

[27] F. Steemers, W. Chang, G. Lee, D. Barker, R. Shen, K. Gunderson, *et al.*, "Whole-genome genotyping with the single-base extension assay.," *Nature methods*, vol. 3, no. 1, p. 31, 2006.

[28] H. Willenbrock and J. Fridlyand, "A comparison study: applying segmentation to array cgh data for downstream analyses," *Bioinformatics*, vol. 21, no. 22, pp. 4084–4091, 2005.

[29] W. Lai, M. Johnson, R. Kucherlapati, and P. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data," *Bioinformatics*, vol. 21, no. 19, pp. 3763–3770, 2005.

[30] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based dna copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.

[31] E. Venkatraman and A. Olshen, "A faster circular binary segmentation algorithm for the analysis of array cgh data," *Bioinformatics*, vol. 23, no. 6, pp. 657–663, 2007.

[32] G. Marenne, B. Rodríguez-Santiago, M. Closas, L. Pérez-Jurado, N. Rothman, D. Rico, G. Pita, D. Pisano, M. Kogevinas, D. Silverman, *et al.*, "Assessment of copy number variation using the illumina infinium 1m snp-array: a comparison of methodological approaches in the spanish bladder cancer/epicuro study," *Human mutation*, vol. 32, no. 2, pp. 240–248, 2011.

[33] D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. Lionel, B. Thiruvahindrapuram, J. MacDonald, R. Mills, *et al.*, "Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants," *Nature biotechnology*, vol. 29, no. 6, pp. 512–520, 2011.

[34] E. Cho, J. Tchinda, J. Freeman, Y. Chung, W. Cai, and C. Lee, "Array-based comparative genomic hybridization and copy number variation in cancer research," *Cytogenetic and genome research*, vol. 115, no. 3-4, pp. 262–272, 2006.

[35] P. Breheny, P. Chalise, A. Batzler, L. Wang, and B. Fridley, "Genetic association studies of copy-number variation: Should assignment of copy number states precede testing?," *PloS one*, vol. 7, no. 4, p. e34262, 2012.

[36] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nature Reviews Genetics*, vol. 2, no. 2, pp. 91–99, 2001.

[37] S. Lin, J. A. Rogers, and J. C. Hsu, "A confidence-set approach for finding tightly linked genomic regions," *The American Journal of Human Genetics*, vol. 68, no. 5, pp. 1219–1228, 2001.

[38] J. Hoh and J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes," *Nature Reviews Genetics*, vol. 4, no. 9, pp. 701–709, 2003.

[39] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[40] D. Zaykin, L. Zhivotovsky, P. Westfall, and B. Weir, "Truncated product method for combining p-values," *Genetic epidemiology*, vol. 22, no. 2, pp. 170–185, 2002.

[41] F. Dudbridge and B. P. Koeleman, "Rank truncated product of p-values, with application to genomewide association scans," *Genetic epidemiology*, vol. 25, no. 4, pp. 360–366, 2003.

[42] H.-C. Yang, C.-Y. Lin, and C. S. Fann, "A sliding-window weighted linkage disequilibrium test," *Genetic epidemiology*, vol. 30, no. 6, pp. 531–545, 2006.

[43] S. R. A. Fisher, S. Genetiker, R. A. Fisher, S. Genetician, G. Britain, R. A. Fisher, and S. Généticien, *Statistical methods for research workers*, vol. 14. Oliver and Boyd Edinburgh, 1970.

[44] E. S. Edgington, "An additive method for combining probability values from independent experiments," *The Journal of Psychology*, vol. 80, no. 2, pp. 351–363, 1972.

[45] S. S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*, vol. 279. Wiley. com, 1993.

[46] B. Wilkinson, "A statistical consideration in psychological research.," *Psychological Bulletin*, vol. 48, no. 2, p. 156, 1951.

[47] F. Dudbridge and B. P. Koeleman, "Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies," *The American Journal of Human Genetics*, vol. 75, no. 3, pp. 424–435, 2004.

[48] R. Fisher, *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[49] S. Stouffer, E. Suchman, and L. DeVinney, "Star, sa, & williams, rm, jr.(1949)," *The American Soldier*, vol. 1, 1949.

[50] R. Littell and J. Folks, "Asymptotic optimality of fisher's method of combining independent tests," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 802–806, 1971.

[51] L. Hedges, I. Olkin, M. Statistiker, I. Olkin, and I. Olkin, "Statistical methods for meta-analysis," 1985.

[52] W. Felier, "An introduction to probability theory and its applications, vol. 1," *IWiley, New York*, 1968.

[53] I. Goods, "On the weighted combination of significance tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 264–265, 1955.

[54] B. F. Manly, *Randomization, bootstrap and Monte Carlo methods in biology*, vol. 70. CRC Press, 2007.

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

[56] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[57] D. Lin, "An efficient monte carlo approach to assessing statistical significance in genomic studies," *Bioinformatics*, vol. 21, no. 6, pp. 781–787, 2005.

[58] Y. Ge, S. Dudoit, and T. P. Speed, "Resampling-based multiple testing for microarray data analysis," *Test*, vol. 12, no. 1, pp. 1–77, 2003.

[59] S. Seaman and B. Müller-Myhsok, "Rapid simulation of< i> p</i> values for product methods and multiple-testing adjustment in association studies," *The American Journal of Human Genetics*, vol. 76, no. 3, pp. 399–408, 2005.

[60] L. Li, B. Fridley, K. Kalari, G. Jenkins, A. Batzler, S. Safgren, M. Hildebrandt, M. Ames, D. Schaid, and L. Wang, "Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression," *Cancer research*, vol. 68, no. 17, pp. 7050–7058, 2008.

[61] M. Davidian, *Nonlinear models for repeated measurement data*, vol. 62. CRC Press, 1995.

[62] L. Li, B. L. Fridley, K. Kalari, G. Jenkins, A. Batzler, R. M. Weinshilboum, and L. Wang, "Gemcitabine and arabinosylcytosin pharmacogenomics: genome-wide association and drug response biomarkers," *PLoS One*, vol. 4, no. 11, p. e7765, 2009.

[63] N. Niu, Y. Qin, B. L. Fridley, J. Hou, K. R. Kalari, M. Zhu, T.-Y. Wu, G. D. Jenkins, A. Batzler, and L. Wang, "Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines," *Genome research*, vol. 20, no. 11, pp. 1482–1492, 2010.

[64] C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and M. E. Hurles, "A robust statistical method for case-control association testing with copy number variation," *Nature genetics*, vol. 40, no. 10, pp. 1245–1252, 2008.

[65] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[66] J. T. Glessner, J. P. Bradfield, K. Wang, N. Takahashi, H. Zhang, P. M. Sleiman, F. D. Mentch, C. E. Kim, C. Hou, K. A. Thomas, *et al.*, "A genome-wide study reveals copy number variants exclusive to childhood obesity cases," *The American Journal of Human Genetics*, vol. 87, no. 5, pp. 661–666, 2010.

[67] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.

[68] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[69] R. J. Boik, "Spectral models for covariance matrices," *Biometrika*, vol. 89, no. 1, pp. 159–182, 2002.

[70] T. Y. Chiu, T. Leonard, and K.-W. Tsui, "The matrix-logarithmic covariance model," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 198–210, 1996.

[71] P. J. Diggle and A. P. Verbyla, "Nonparametric estimation of covariance structure in longitudinal data," *Biometrics*, pp. 401–415, 1998.

[72] T. Leonard and J. S. Hsu, "Bayesian inference for a covariance matrix," *The Annals of Statistics*, pp. 1669–1696, 1992.

[73] R. Yang and J. O. Berger, "Estimation of a covariance matrix using the reference prior," *The Annals of Statistics*, pp. 1195–1211, 1994.

[74] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

Vita

Yinglei Li

University of Kentucky Department of Statistics

Education

Master of Science in Statistics, University of Kentucky, 2010

Master of Economics and Management in Technical Economics and Management, Tongji University, China, 2008

Bachelor of Science in Textile Trading, Donghua University, China, 2003

Employment

Research/Teaching Assistant *Aug 2008 to Aug 2014*

Statistics Department, University of Kentucky

Graduate Research Assistant *May 2005 to Aug 2007*

Tongji Investment Institution, Tongji University

Human Resource Assistant *Aug 2003 to Jul 2004*

Golden Ideas consulting Co., Shanghai, China

Selected Publications

1. Yinglei Li and Patrick Breheny (2013). Kernel-Based Aggregation of Marker-Level Genetic Association Tests Involving Copy-Number Variation. *Microarrays*, 2, pp265-283