Principal Component Analysis in High Dimensional Data: Application for Genomewide Association Studies

by
Seunggeun Lee

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2010

Approved by:

Fei Zou, Advisor

Fred A. Wright, Advisor

Yufeng Liu, Reader

Patrick A. Sullivan, Reader

Hongtu Zhu, Reader

Abstract

SEUNGGEUN LEE: Principal Component Analysis in High Dimensional Data: Application for Genomewide Association Studies.
(Under the direction of Fei Zou and Fred A. Wright.)

In genomewide association studies (GWAS), population stratification (PS) is a major confounding factor which causes spurious associations by inflating test statistics. PS refers to differences in allele frequencies by disease status due to systematic differences in ancestry, rather than causal association of genes with disease. PCA is commonly used to infer population structure by computing PC scores, which are subsequently used for control of population stratification.

Even though PCA is now widely used for PS adjustment, there are still challenges for PCA based effective PS control. One common feature of the genomic data is the strong local correlation among adjacent loci/markers caused by linkage disequilibrium (LD). It is known that this local correlation can have a negative effect on estimated PC scores and produce spurious PCs which do not truly reflect underlying population structure. To address this problem, we have employed a shrinkage PCA approach where coefficients are used to down-weight the contribution of highly correlated SNPs in PCA.

Another challenge in PC analysis is choosing which PCs to include as covariates to adjust population stratification. While searching for a reasonable measure for PC selection, we have found the precise relationship between genotype principal components and inflation of association test statistics. Based on this fact, We propose a new approach, called EigenCorr, which selects principal components based on both their eigenvalues and their correlation with the (disease) phenotype. Our approach tends to select fewer principal components for stratification control than does testing

of eigenvalues alone, providing substantial computational savings and improvements in power.

Under many circumstances, it is of interest to predict PC scores. Although PC score prediction is commonly used in practice, characteristics of the predicted PC scores have not been systematically studied. Under high dimensional settings we have found that the naïve predicted PC scores are systematically biased toward 0, and this phenomenon is largely due to the inconsistency of the sample eigenvalues and eigenvectors. We have extended existing convergence results of sample eigenvalues and eigenvectors and derived asymptotic shrinkage factors. Based on these asymptotic results, we propose the bias-adjusted PC score prediction.

Acknowledgments

I would like to express my deepest gratitude to Dr. Fei Zou and Dr. Fred A. Wright, my advisors, for their support and guidance throughout the research. I would also like to thank my committee members Dr. Patrick A. Sullivan, Dr. Hongtu Zhu and Dr. Yufeung Liu, for their very helpful insights, comments and suggestions. I would like to dedicate this dissertation to my fiancee Sehee, father Jehong Lee, mother Insook Lim, and two sisters Juhee Lee and Minkyong Lee, and all of my friends for their endless love, encouragement, and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Overview

With development of single nucleotide polymorphism (SNP) microarray technology, genome-wide association studies (GWAS) have emerged as an effective tool to unravel genetic components of complex diseases. GWAS find associated genes by examining the allele frequencies of SNPs across the disease phenotype. However, if there are systematic differences of allele frequencies among sub-populations, and each sub-population has different disease prevalence, this confounding effect of population stratification (PS) can cause false positive associations by inflating test statistics. There have been considerable efforts to address PS. The method of genomic control (Devlin and Roeder, 1999; Devlin, Roeder and Wasserman, 2001) was among the first attempts to address this problem. It directly estimates the inflation of test statistics by computing the median of chi-square test statistics and adjusts inflation of test statistics simply by dividing the amount of estimated inflation. For this purpose, genomic control assumes that the inflation of test statistics is constant for all SNPs, and estimates the inflation factor from null SNPs, which are assumed not to be associated with the disease phenotype. Genomic control works well in candidate gene studies; however, it does not produce satisfactory results for GWAS because the constant inflation assumption is too strong to be satisfied (Freedman et al., 2004; Marchini et al., 2004; Devlin, Bacanu and Roeder,

2004).

Another approach for addressing PS is through principal component analysis (PCA) (Jolliffe, 2002), which was proposed by Price et al. (2006), and it has become very popular. PCA directly estimates an ancestry of each sample by computing PC scores from all SNPs, and adjusts PS using PC scores as ethnicity covariates. PCA has been applied in many GWAS and has successfully adjusted PS. In several studies, PC scores often reflect already known genetic structure (Price et al., 2008; Tian et al., 2008). One great advantage of PCA is that it can utilize all SNP genotypes, and thus it can detect even very subtle PS.

Even though PCA is now widely used for PS adjustment, there are still challenges for PCA based effective PS control. First of all, current PCA is not designed to address the linkage disequilibrium. Linkage disequilibrium, which causes strong local correlations among adjacent SNPs, can affect negatively estimated PC scores. Also, it can produce spurious PCs which do not truly reflect underlying population structure (Fellay et al., 2007). For example, it has been observed that chromosome inverted regions of Chr8 and Chr17 drived highly ranked PCs among European samples. To address this problem, we have employed a weighted PCA approach (Greenacre, 1984) where coefficients are used to down-weight the contribution of highly correlated SNPs. In simulation and real data analysis, our Shrinkage PCA has successfully adjusted for LD and has been shown to have better performance than competing population stratification control methods such as SNP pruning and regression based method (Patterson, Price and Reich, 2006).

In the second topic, we develop a method for PC selection. After computing PCs, we have PCs as many as the number of samples. Apparently, we cannot use all of those PCs as covariate, and thus PC selection is necessary. Price et al. (2006) originally suggested using the 10 PCs with the highest eigenvalues. Later, Patterson, Price and Reich (2006) have proposed using the Tracy-Widom (TW) statistic (Johnstone, 2001)

to assess statistical significance of eigenvalues in order to select PCs. These approaches select PCs only based on eigenvalues, as a result, they may select PCs which does not inflate test statistics. In Chapter 3, we first identify the precise relationship between PCs to the inflation of test statistics, and it appears that the inflation of test statistics is a joint function of eigenvalues and correlation coefficient between PC and outcome phenotype. From this fact, we propose the *EigenCorr* method, which selects PCs according to the compound score of the corresponding eigenvalue and correlation. In simulation and real data analysis, we show that *EigenCorr* chooses a much smaller number of PCs than does the Tracy-Widom test, while successfully controlling type I error. As a result, *EigenCorr* reduces computation time substantially but has higher statistical power and improves type I error control. In addition to these practical advantages, our research has established the connection of PCs to the spurious inflation of test statistics, thus providing a theoretical justification for employing PCA in PS adjustment.

My third topic considers the PC score prediction. Under many circumstances, it is of interest to predict PC scores. For GWAS, it is known that PC analysis with related subjects tends to generate spurious PC scores which do not reflect the true underlying population substructures. To overcome this problem, it is common in practice to exclude some related samples and only apply the PC analysis to those remaining unrelated samples. To use those excluded samples in downstream analysis, we need to predict PC scores. The standard method to predict PC scores is multiplying the data vector of the new sample with the loading coefficients from the PC analysis. However, we have found that the predicted PC score from this standard method is systematically biased toward 0, and this phenomenon is deeply related to the inconsistency of eigenvectors of the sample covariance matrix to the eigenvalues of the population covariance. The convergence of sample eigenvalues and eigenvectors has been studied under the Random

Matrix context. With spiked model assumption, Baik and Silverstein (2006) has shown the convergence of sample eigenvalues, and Paul (2007) has shown the convergence of sample eigenvectors. Those theoretical results show that sample eigenvalues and eigenvectors are not consistent to the population eigenvalues and eigenvectors. In Chapter 4, we extend those existing theoretical results and derive asymptotic shrinkage factors of predicted PC scores. Based on this theoretical result, we propose bias adjusted PC scores. From numerical studies and real data analysis, we show that the shrinkage bias appears in real data, and our approach can successfully adjust it.

Chapter 2

Control of Population Stratification Using Correlated SNPs by Shrinkage Principal Components

Association studies using unrelated individuals have become the most popular design for mapping complex traits. Among the major challenges of association mapping is avoiding spurious association due to population stratification. Several approaches have been proposed to handle population stratification using marker genotypes, including genomic control, structured assessment of ancestry, principal component analysis and partial least squares analysis of phenotype and genotypes. Among these approaches, only genomic control and principal components can handle the high-dimensional data encountered in genome-wide association studies. Empirical studies favor the principal components approach for its power and error control properties. All of the stratification-control methods impicitly assume that the markers are in linkage equilibrium, a condition that is rarely satisfied in genome scans. Moreover, the impact of linkage disequilibrium on these methods for stratification control has not been carefully articulated or examined. In this paper, we extend the principal components approach to all available

markers, regardless of the linkage disequilibrium patterns. We illustrate the behavior of our approach using simulated and real data, and several practical issues are discussed.

2.1 Introduction

Over the past two decades, considerable effort has been expended to detect and map the genetic loci contributing to complex diseases. Association and linkage studies are the two main strategies for this purpose. Association studies using unrelated individuals have become the dominant study design for genome-wise association scans (GWAS), partly because accrual of patients and controls is easier than for family-based designs. It has been argued that direct association mapping is more powerful than linkage analysis for identifying loci with small effects (Risch and Merikangas, 1996). Association mapping is typically also more precise, because the assocation of genotypes with disease drops rapidly in the vicinity of a risk locus, due to a large number of historical recombinations for an ancient variant.(Cardon and Bell, 2001; Cardon and Palmer, 2003; Daly and Day, 2001; Elston, 1998; Schulze, McMahon and Methods, 2002). Several successful GWA studies have been reported recently, identifying genetic variants contributing to Type 2 diabetes (Saxena et al., 2007; Scott et al., 2007; Sladek et al., 2007; Zeggini et al., 2007), breast cancer (Easton et al., 2007), and numerous other diseases. However, it has long been discussed that association studies are susceptible to underlying population stratification, which can produce spurious association (Cardon and Palmer, 2003). A number of techniques have been proposed to account for population substructure in designs using unrelated individuals. These techniques include using aggregate summaries of association statistics to estimate the inflation produced by stratification (*genomic control* of Devlin and Roeder (1999); Schork et al. (2001)). Other approaches use marker genotypes to model the population structure directly, performing association tests conditional on the inferred structure (*structured assessment* of Pritchard,

Stephens and Donnelly (2000), Satten, Flanders and Yang (2001) and Zhu et al. (2002) developed similar approaches which account for uncertainty in stratum classification. Similarly, Zhang, Zhu and Zhao (2003) have proposed to use principal component analysis (PCA) to estimate genetic background covariates, which then are used in adjusted tests of association. One limitation of the classical PCA methods is that the number of markers cannot exceed the number of subjects. Price et al. (2006) exploited the structure of rescaled genotype matrices to extend the PCA method to modern genome scans, in which hundreds of thousands of SNPs are genotyped. Due to the popularity of this approach (implemented in the software Eigensoft), we will refer to it as the standard PCA approach.

A number of investigators have considered the number of markers required to identify and control for population stratification. Earlier efforts primarily envisioned stratification at the level of continental populations (Bacanu, Devlin and Roeder, 2000; Devlin and Roeder, 1999; Pritchard, Stephens and Donnelly, 2000), requiring as few as 20-500 markers (Pritchard and Rosenberg, 1999). However, with so few markers, sensitivity can be poor under moderate stratification (Freedman et al., 2004; Hao et al., 2004). For this reason, modern PCA-based methods are appealing, because they can in principle use the entire dataset for stratification control, ranging from moderate-scale candidate gene studies to whole genome scans.

Unfortunately, the use of all available data presents a problem, as well. Except for genomic control, all of the methods described above assume that the markers used for stratification control are unlinked. Falush, Stephens and Pritchard (2003) proposed a procedure to identify population structure using correlated markers, but their method is limited and not applicable to situations with tightly linked markers. Price et al. (2006) initially suggested that markers in linkage disequilibrium have little effect on PC-based stratification analysis, but subsequently proposed reducing marker

linkage disequilibrium via regression (Patterson, Price and Reich, 2006). Fellay et al. (2007) utilized a thinning approach in which only a subset of markers with low pairwise correlation was retained for stratification control. The use of thinning involves discarding large and potentially informative portions of the data, and identification of the low-correlation subset can involve considerable computation, and perhaps iteration. Although the potential problems posed by dependent markers are increadingly recognized, to our knowledge the consequences of using dependent markers has not been carefully investigated.

In this paper, we demonstrate that LD patterns in genome-wide association datasets can distort the techniques for stratification control, showing subpopulations that reflect localized LD phenomena rather than plausible population structure. Further analysis based on such spurious stratification may provide inadequate protection from genuine stratification, and may reduce mapping power in key regions of the genome. To account for the LD structure, we propose a simple modification of the standard PCA approach to automatically adjust for the correlations among markers and accurately infer population stratification. The usefulness of our approach is demonstrated by simulations and application to candidate gene and genome-wide association studies.

2.2 Materials and Methods

When principal components are used to identify subpopulations, it is implicitly assumed that all variables are of similar importance (Chatfield and Collins, 1981; Morrison, 1976). In association mapping, some groups of SNPs may be highly correlated (both positive and negative) due to localized LD, while other sets of SNPs may have low correlation. PC analysis finds projections of the data with high variability. Correlated SNPs will therefore have high loadings, because correlated random variables can generate linear combinations with high variability. As we demonstrate, the net effect

is to give higher weight to groups of correlated SNPs, although there is little reason to believe that such SNPs will perform well in differentiating among subpopulations. An intermediate goal, therefore, is to eliminate the distorting effect of the redundant information provided by groups of highly correlated SNP genotypes. The weighted PCA method of Greenacre (1984) was proposed for similar problems by using weights or new PCA metrics. In time series applications, Diamantaras and Kung (1996) have used weighted covariance matrices, with weights decreasing geometrically with the distance in time between observations. In atmospheric science, weights have been used to account for uneven spacing between sampling locations (Cheng, 2002). Similar weighting ideas might be used in GWAS analysis, as pronounced linkage disequilibrium is largely a localized phenomenon on the genome. However, the extent of linkage disequilibrium between loci is not a fixed function of physical distance (Maniatis et al., 2002), and varies across subpopulations (Service et al., 2006). The use of data-driven weighting would be preferred, to directly address the problematic effects of correlation in the data at hand. In addition, any weighting scheme must be scalable up to the common GWAS situation in which the number of variables (SNPs) is far larger than the number of observations. Accordingly, we propose a unified shrinkage method that deals with all markers simultaneously, effectively down-weighting SNPs that belong to highly correlated groups, while leaving independent SNPs unchanged.

Our proposed shrinkage method is a modification of the PCA method of Price et al. (2006). Let g_{ij} represent the (i,j)th element of the genotype matrix g, corresponding to SNP i and individual j, $i = 1..., M$ and $j = 1, ..., N$. By convention, g_{ij} is coded numerically as the number of copies of a referent allele (the minor allele, say) for the SNP. Each row i of g is first mean-centered around $\mu_i = \sum_j g_{ij}/N$ (missing entries are excluded from the computation of μ_i and subsequently set to 0). Row i is then scaled by dividing each entry by the standard deviation $\sqrt{p_i(1-p_i)}$, where $p_i = (1+\sum_j g_{ij}))/(2+$

$2N$) is the estimated allele frequency at SNP i. Denoting the resulting matrix X, Price et al. (2006) define the kth axis of variation to be the kth eigenvector of C, where $C = X^T X$. The coordinate j of the kth eigenvector represents the ancestry of individual j along the kth axis of variation. Unlike the classical application of principal components (Jolliffe, 2002) which is based on the $M \times M$ matrix $D = XX^T$, standard PCA for genome-wide studies (Price et al., 2006) uses the $N \times N$matrix C, which is typically of much smaller dimension in GWAS studies. The justification for this approach arises from the close relationship between singular value decomposition and PCA when the latter is performed on mean centered data (see, for example, Wall, Rechtsteiner and Rocha (2003)). EigenSoft employs the singular value decomposition$X = USV^T$, where U is an $M \times N$ matrix whose kth column contains coordinates $u_{ik}$of each SNP i along the kth principal component, S is a diagonal matrix of singular values, and V is an $N \times N$ matrix whose kth column contains ancestries v_{jk} of each individual j along the kth principal component. It follows that $X^T X = VS^2V^T$. Thus, the columns of V are the eigenvectors of the matrix $X^T X$. After PCA analysis, pairwise scatter plots of the top few PC axes are often used to investigate potential population stratification. In addition to the PCs, the loading coefficents associated with each PC can be calculated, but are often overlooked. Loadings calculate the contribution of each SNP for a given PC. When $M \leq N$, the loadings can be uniquely determined; otherwise, they are not. For a given PC k , at SNP i, $u_{ik}$is the loading coefficient for the SVD analysis at the SNP. The loadings can be calculated as $\sum_j v_{jk} x_{ij} / \sqrt{e_k}$, where e_k is the corresponding eigenvalue of the PC k . These loadings are closely related to the gamma coefficients $\gamma_{ik} = \sum_j v_{jk} g_{ij} / \sum_j v_{jk}^2 \cong \sum_j v_{jk} g_{ij}$ described in Price et al. (2006). We have $\sum_j v_{jk} =$

0 and$\sum_j v_{jk}^2 = 1$ when there are no missing genotypes at SNP i, and we therefore have

$$\gamma_{ik} = \sum_j v_{jk} g_{ij} = \sum_j v_{jk} \left(x_{ij} \sqrt{p_i(1-p_i)} + \mu_i \right) = \sqrt{p_i(1-p_i)} \sum_j v_{jk} x_{ij} - \mu_i \sum_j v_{jk}$$
$$= u_{ik} \sqrt{e_k p_i (1-p_i)}$$

If some genotypes are missing at SNP i, the above equality remains approximately correct, unless the rate of missing genotypes is high.

Eigensoft treats each SNP in an equal manner. However, as we demonstrate below, the direct use of C in fact results in loadings that can be dominated by small groups of correlated SNPs. To correct for this phenomenon, we propose a new approach to weighted PC analysis. First, we define an M-vector w of SNP weights, and accompanying diagonal matrix W with weights w on the main diagonal. Then we create a new $M \times N$ matrix $\tilde{X} = WX$, which is directly substituted for X in the PC analysis as described in Price et al. (2006).

Our choice of weights follows the logic that linear combinations of genotypes (which comprise the eigenvectors) should exert influence determined by the amount of independent information. We heuristically choose weights $w_i = 1 \Big/ \sqrt{1 + \sum_{i' \neq i} r_{ii'}^2}$ for SNP i, where $r_{ii'}^2$ is the observed squared Pearson correlation between ith and i'th SNPs. In practice, our summation over SNPs i' in calculating the weights is performed only in the vicinity of i, in order to filter out the cumulative effect of random apparent correlation across the genome. We will refer to the set of such SNPs as $window[i]$, and these SNPs may range up to several hundred kb from SNP i, as chosen by the researcher and appropriate to the platform. In addition, the effects of noise in the use of r^2 (which must always be positive) is reduced by requiring that r^2 exceed a threshold c. Thus the precise weighting scheme is $w_i = 1 \Big/ \sqrt{1 + \sum_{i' \neq i, i \in window[i]} r_{ii'}^2 I[r_{ii'}^2 > c]}$. For this paper we use c=0.2, but other choices are possible and remain under investigation.

In this manner SNPs that are highly correlated with each other are down-weighted, de-emphasizing their importance. Our choice of weights has the following desirable characteristics.

If all markers are independent and there exists no population stratification, $r^2_{ii'} \cong 0$ for all i' and therefore $\tilde{X} \cong X$. If all pairs of $m0$ markers have $r^2 = 1$ with each other and zero correlation with other markers, then the weighting factor is $1/\sqrt{m_0}$, effectively providing variance contributions of the $m0$ markers equivalent to that of a single marker. Finally, if correlation among all pairs of markers is nonzero but approximately equal, as would be produced in idealized models of population stratification, then the weights will also be constant. Therefore $\tilde{X} \cong cX$ for some c, and the net effect is that markers are treated equally, as in standard PCA.

Plots of loading coefficients display the contribution of each SNP to a given PC, but also present a global picture of the influence of SNPs in regions of high LD. Our experience suggests that routinely checking plots of loading coefficients is very useful in identifying regions with high influence on a PC.

2.3 Simulation Studies and Applications

We first show how LD affects PC analysis based on simulations of an association study with 200 markers, such as might be performed in a candidate gene association study or a follow-up to a GWAS study. We then show a simulated GWAS study with 100K SNPs. In addition, we apply our proposed method to a real candidate gene association study and to a GWAS study. Finally, we investigate the power and type I error issues from the downstream analysis after the PCA analysis by simulated GWAS data.

Simulation 1 (candidate SNP analysis, independent markers): A stratified population with two sub-populations was simulated. A total of 400 individuals were sampled, with 200 from each sub-population. 200 markers were simulated, each with 3

possible genotypes and minor allele frequency ranging uniformly from 0 to 0.5. All markers were unlinked and in linkage equilibrium within each sub-population. We applied both the standard PCA approach and our shrinkage PCA to the dataset. Scatter plots of the top two PCs are presented in Figure 2.1. Clearly, when markers are in linkage equilibrium, both PCA methods give similar results.

Simulation 2 (candidate SNP analysis, markers with varying correlation): Again a stratified population was simulated, with the same number of individuals and markers as in simulation 1. However, here two subsets of markers were chosen to be in high LD with each other within each sub-population. The results of simulation 2 (some markers in LD) are presented in Figure 2.2. Under the standard PCA (Figure 2.2 left panels), the data points form groups that are mainly influenced by the SNPs in high LD. In this manner, subjects may be misclassified, or unnecessary extra stratification performed. Examination of the loadings for the first two PCs shows that they are dominated by the blocks of markers in high LD. The shrinkage approach (Figure 2.2 right panels), in contrast, retrieves the original sub-populations successfully. Examination of the loadings for the shrinkage PCA shows that the SNPs in the LD blocks have been downweighted considerably.

Simulation 3 (GWAS data based on Hapmap Samples): We conducted GWAS simulation to investigate the performance of the PCA methods on substantially stratified populations. With *HapSample* software, we first simulated 450 CEU samples, 50 YRI samples, and 50 JP+CH samples respectively using the SNPs on the Affymetrix 100K array (Wright et al., 2007). *HapSample* generates data by resampling from existing phased Hapmap datasets and therefore preserves the observed local LD structure in Hapmap samples. We then generated additional 225 individuals with mixed genomes from the three populations, using our modified codes from *HapSample*. Specifically, we generated 50 admixture samples of CEU and YRI, 50 admixture samples of CEU and

JP+CH, and the reminder125 are admixture samples of the three populations. That is, for the ith admixture sample, we have

$$(p_{i1}, \quad p_{i2}, \quad p_{i3}) = \begin{cases} (u, \quad 1-u, \quad 0), & i \leq 50 \\ (u, \quad 0, \quad 1-u), & 51 \leq i \leq 100 \\ (g_1, \quad g_2, \quad 1-g_1-g_2), & 101 \leq i \leq 225 \end{cases}$$

where(p_{i1}, p_{i2}, p_{i3}) are the corresponding CEU, YRI and JP+CH genome proportions, respectively, $u \sim Unif(0,1)$ and $(g_1, g_2) \sim Dirichlet(70, 15, 15)$. The final simulated data has 775 samples and 109,723 SNPs. Figure 2.3 presents the scatter plots of the top two PCs derived from the following four different methods: 1) Standard PCA with no LD correction, 2) Shrinkage PCA, 3) Regression PCA (Patterson, Price and Reich, 2006) in EigenSoft, and 4) Thinning PCA in plink which is based on pairwise correlation (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker and Daly, 2007). For the regression PCA, we followed the recommendation of EigenSoft, where previous 2 SNPs were used in the regression analysis. For the thinning PCA, we follow the setting "–indep-pairwise 1500 150 0.2" in Fellay et al. (2007), which ended up with 49,823 SNPs for the PCA analysis. Clearly, for substantially stratified populations where a large number of SNPs are available, all four methods perform equally well.

Real data analysis 1 (candidate gene modifier study of Cystic Fibrosis): Here we use a real example dataset from a candidate gene modifier study of Cystic Fibrosis (CF) underway at the University of North Carolina and Case-Western Reserve University. Over 1000 SNPs have been genotyped in 263 severe CF patients and 545 mild CF patients, using the Illumina 1536 platform. Among these SNPs, 81 were autosomal ancestry-informative markers (AIMs), chosen as the most informative SNPs (in terms of allele frequencies) from a list of 200+ potential AIMs provided by Illumina, Inc. in

2006. These AIMs were genotyped for the express purpose of controlling population stratification for the remaining candidate SNPs. Among the 808 patients, 782 were self-reported Caucasians, 14 were Hispanic, 5 were African-American and the remaining 7 reported as belonging to other ethnicity groups.

The genotyped AIMs were carefully selected, with known high *Fst* values between the Caucasians and West African popualtions. At the time, the effect of LD on population stratification control was not explicitly considered, and several sets of SNPs exhibited appreciable correlation (2 SNPs on chromosome 1, 2 SNPs on chromosome 7 and 3 SNPs on chromsome 3). Standard PCA analysis (Figure 2.4, left panels) shows that PCA analysis is highly influenced by the high LD SNPs, and similar results were observed for STRUCTURE analysis (data not shown). The right panels of Figure 2.4 shows the PC results from shrinkage PCA, which is much less sensitive to the LD among SNPs. The SNPs with high LD have loadings of large magnitude for PC1 in traditional PCA analysis, while the shrinkage PCA analysis eliminates this artificial effect. The results of our proposed method are clearly superior – the African-American and Hispanic subjects are more clearly distinguished from Caucasians on PC1 (Figure 2.4). Interestingly, one of the subjects labeled as Caucasian (indicated by arrow in panel B), was flagged as an outlier by our shrinkage PC analysis, but not by standard PCA. A subsequent check of the recruitment database revealed a data entry error, and the subject was in fact a self-reported African-American. This example shows the utility of the shrinkage PCA approach in candidate gene studies, in which perhaps several hundred SNPs are genotyped. Despite the considerable attention generated by GWAS in recent years, we anticipate that smaller scale candidate gene studies will remain popular, due to cost considerations, or as follow-up studies to confirm results from genome scans.

Real data analysis 2 (Hapmap CEU amd TSI data) As shown by our Simulation 2, substantial population stratification can be easily detected with any of the

existing PCA methods. An important question is how those methods perform for subtle population stratification. Below we address this issue by the Phase 3 CEU and TSI Hapmap unrelated samples. The Plink formatted data was downloaded from the Hapmap website (http://ftp.hapmap.org/phase_3/?N=D). We removed all children whose parents are also Hapmap samples. Additionally we excluded one CEU subject who has a very high estimated identical by descent (IBD) value (> 0.8) with another CEU sample. The final dataset after the filtering contains total 185 samples (108 CEU and 77 TSI samples respectively). The CEU samples are known to have the northern and western European ancestry, while the TSI samples are Toscans from Italy. Therefore the two groups represent the Northern-Western and Southern Europeans, respectively. We restrict our analysis on SNPs from one chromosome (which is Chromosome 15 for this example) as done in (Miclaus, Wolfinger and Czika, 2009) for further comparison between our shrinkage PCA and other three existing LD correction methods described in Simulation 3 Section on their abilities in detecting subtle population stratifications. SNPs with missing rate bigger than 0.1 or MAF less than 0.01 were excluded, resulting in 38,711 SNPs. Again, for the regression PCA, we followed the EigenSoft recommendation. For the thinning PCA, we used the setting in (Fellay et al., 2007) which resulted in 3,218 SNPs. Figure 2.5 shows the scatter plots of the top 2 PCs of all four methods. Clearly, our shrinkage PCA outperforms the other three methods on differentiating the two groups. Figure 2.6 compares the ROC curves of using the 1st PC to classify the two sub-populations. Again, our shrinkage PCA beats the other three methods. In addition, we tested the Hardy Weinberg Disequilibrium SNP selection method (Miclaus, Wolfinger and Czika, 2009), where SNPs were selected if their associated p-value on Hardy Weinberg Equilibrium test is less than 0.01. Only 254 SNPs were selected based on this criterion, and the top two PCs derived from these 254 SNPs showed poor performance due to these limited number of SNPs.

Real data analysis 3 (GWAS study of Schizophrenia): A third real dataset is from a GWAS study of schizophrenia, obtained from the GAIN consortium. The filtered version of the corresponding General Research Use (GRU) dataset consisted of 2,601 individuals of European ancestry with 729,454 SNPs, and was downloaded from the dbGap database [Version 2, Accession number: phs000021.v2.p1]. We filtered out highly related or duplicated samples, and markers with a high missingness rate ($>$ 5%) or a low minor allele frequency (< 0.01). For simplicity, sex chromosome markers were excluded, and the final data set used for stratification analysis had 2,559 samples with genotypes (1152 cases, 1368 controls and 39 with missing case-control status) and 701,859 SNPs. For calculation of weights wi, the shrinkage PC method used 300 SNPs in the vicinity of each i as the window, and c=0.2.

Scatter plots of the top 2 PCs from the two stratification analysis methods are presented in Figure 2.7. Standard PCA analysis using the original data provides results with major groups that are almost certainly spurious. After the shrinkage PCA approach is applied, the result appears similar to previous analyses of populations with mixed European ancestry (e.g., Figure 2 in Price et al. (2006)). Plots of loading coefficients for these analyses are given in Figure 2.7. The top 4 PCs from standard PCA are highly influenced by a few genomic regions. The lactase gene region on 2q21-2q22 is highly influential for PC1, which is consistent with a northern-southern cline in haplotype frequencies (Hollox et al., 2001). Interestingly, our shrinkage PCA preserves this feature, and the correlation of PC1 from standard PCA and that of shrinkage PCA is 0.98. However, regions with high loadings on PC2 (8p23), PC3 (2q21, 6p21-22, 17q21) and PC4 (6p21-22) from standard PCA have all disappeared after the shrinkage PCA, suggesting that the high impact of those regions (except for lactase, captured in PC1) is simply due to high regional LD. The regions 8p23 and 17q21 coincide with

two previously reported common inversions in European populations (, N.d.; Stefansson et al., 2005). The chromosome 8 inversion region has been similarly reported by Fellay et al. (2007) in their GWAS study of HIV-1. These inversions have only been discovered in the last several years, and it is in many ways remarkable that they be detected so readily using GWAS genotypes. Presumably the LD is maintained by selection against crossovers in such regions, but not necessarily indicative of ancestry if well-mixed within the population. The 6p21 region coincides with the MHC region, for which extensive linkage desiquilibrium has been described (de Bakker et al., 2006). We conclude that the shrinkage PCA approach provides appropriate downweighting, so as not to be unduly influenced by such regions, while retaining the influence of SNPs and regions indicative of true stratification.

Type I and Power issues: In this simulation, a stratified population with two subpopulations was simulated, with 1800 samples from population one and 200 samples from population two. 50,000 independent markers were first simulated, with minor allele frequency ranging uniformly from 0.05 to 0.5. In order to mimic realistic stratification, *Fst* values were simulated by drawing from the density$0.99\times\chi^2_1/0.03^2+0.01\times U(0,0.05)$, which is a mixture of a scaled chi-square distribution and the uniform distribution on [0, 0.05]. Previous studies have shown that even within European populations, SNPs with *Fst* values ranging from 0.2 to 0.3 between northern and southeastern subpopulations can be observed (Bauchet et al., 2007). Accordingly, we augmented the original *Fst* values with an additional 20 SNPs with high *Fst* values uniformly distributed on [0.1, 0.3]. The minor allele frequencies of the sub-populations were simulated using the Balding-Nichols model (Balding and Nichols, 1995*a*). To create LD blocks, we randomly picked 2500 seed SNPs from the simulated 50,000 SNPs for constructing the blocks. For each seed marker, an additional 20 SNPs were simulated, with correlations ranging from 0.75 to 0.85 with the seed marker, resulting in a total of 100,000 SNPs. The

net result was that half of the markers resided in highly correlated blocks and half consisted of independent markers. Note that the simulation setup described here is relatively favorable to standard PCA, as we did not incorporate extreme blocks of very highly correlated markers that can dominate standard PCA analysis (illustrated in the real data analyses below).

We applied both standard PCA and the proposed shrinkage PCA to the simulated data. Scatter plots of the top two PCs from the two methods are presented in Figure 2.9. Clearly, standard PCA lacks power to identify the two sub-populations, while the shrinkage PCs differentiate the two sub-populations successfully. To investigate if the two methods properly control type I error, we simulated a case-control outcome variable which was related to the sub-populations. Using z to denote the population (z=0 for population one, z=1 for population two), we simulated the data using $\log \frac{P(Case|z)}{P(Control|z)} = 2.5\ z$. In other words, the sub-population status and case/control status were related with a log odds ratio of 2.5, and the resulting datasets had average 1084 cases, 916 controls across the simulations. However, the case/control status was independent of any SNP genotypes within each sub-population. To save computational time, we computed the p-values of only the 20 SNPs with highest *Fst*, reasoning that these SNPs make the greatest contribution to inflated Type I error. In this manner, by applying genome-wide appropriate thresholds to these SNPs for each of 1000 simulations, we obtained a lower bound for the overall Type I error. We emphasize that the actual *Fst* values cannot be known to the researcher without knowledge of the subpopulation indices, and so stratification control is an essential part of the analytic process.

The results of the simulations are given in Table 2.1. Clearly, the standard PCA approach does not control the type I error properly. For 100,000 markers, even if conservative Bonferroni family-wise error (FWER) thresholds are intended, the true FWER is much higher. For example, p-value thresholds of $1X10^{-6}$ and $5X10^{-7}$provide

intended FWER values of no greater than 0.10 and 0.05, repectively. However, Table 1 shows that the true Type I errors are at least 0.279 and 0.195 for this setup. In contrast, our shrinkage method seems to control the Type I error adequately, since the top 20 SNPs with the highest *Fst* have a negligible effect on the Type I error.

2.4 Discussion

The PCA approach can capture both subtle and extensive variation due to both genomic and experimental features. With the availability of $> 10^5$ genetic markers, self-reported race may no longer required as a proxy for ancestry. The principal components method is computationally efficient and uses the genotype matrix to infer continuous axes of genetic variation (eigenvectors) which then serve as covariates in the down-stream analysis. This method is widely used in GWAS studies to robustly control for stratification effects, while preserving statistical power. However, PCA is highly influenced by sets of SNPs with high LD. Using SNPs with high LD for PCA may distorts population substructures, which is more true for data with subtle population stratification. To our knowledge, this paper appears to be the first that carefully investigates LD structure on PC analysis. Our shrinkage PCA approach has been shown to effectively remove the artifactual effect of correlated SNPs, and so can successfully recovers underlying population structure that is not apparent from standard PCA. The proposed method is essentially a standard PCA approach on a shrunken genotype data, and much easier to implement than other approaches, such as the regression based PCA of Patterson, Price and Reich (2006).

Groups of SNPs in high LD may have an even greater effect on candidate gene studies than on GWAS studies. Although GWAS studies are becoming a primary design for studying complex traits, candidate studies remain important, and are often employed for replication and validation. In this setting, a set of ancestry-informative

markers is typically used, and our approach applies equally well as with GWAS studies.

We note that the shrinkage method intends to remove only the effects of local LD, as subtle long-range LD (for example, across chromosomes) reflects true population sub-structure, and our weighting scheme leaves the effects of long-range LD intact. However, other weighting schemes are possible, and the most efficient weighting scheme for elucidating population structure remains unknown and of great interest. Also, we point out that substructure inference is not simply a matter of an error control, as other types of procedures (such as genotype imputation at untyped SNPs) can depend on accurate ancestry inference.

2.5 Tables and Figures

Table 2.1: Rejection frequency of 20 high F_{st} SNPs under null genetic association

p-value threshold	Expected # of rejections	No Adjustment	Known Strata	Standard PCA	Shrinkage PCA	Thinning PCA	Regression PCA
10^{-5}	2	10000	2	618	24	303	111
10^{-6}	0.2	10000	0	142	5	62	21
5×10^{-7}	0.1	10000	0	95	1	34	11
10^{-7}	0.02	10000	0	29	0	12	0

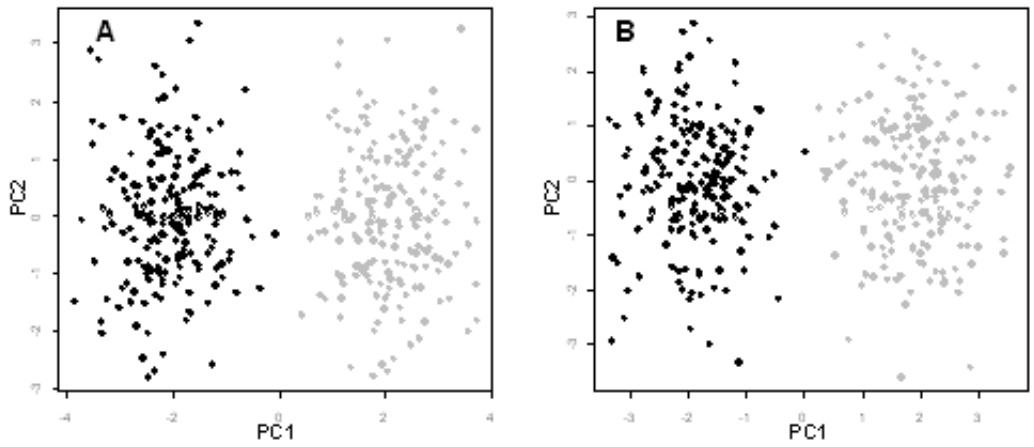

Figure 2.1: Simulation 1 (independent markers). A stratified population with all SNPs independent within each subpopulation. 200 markers for 400 individuals were simulated as described in the text. The different subpopulations are indicated in gray and black. Both standard PCA (left panel) and shrinkage PCA (right panel) effectively separate individuals according to subpopulation.

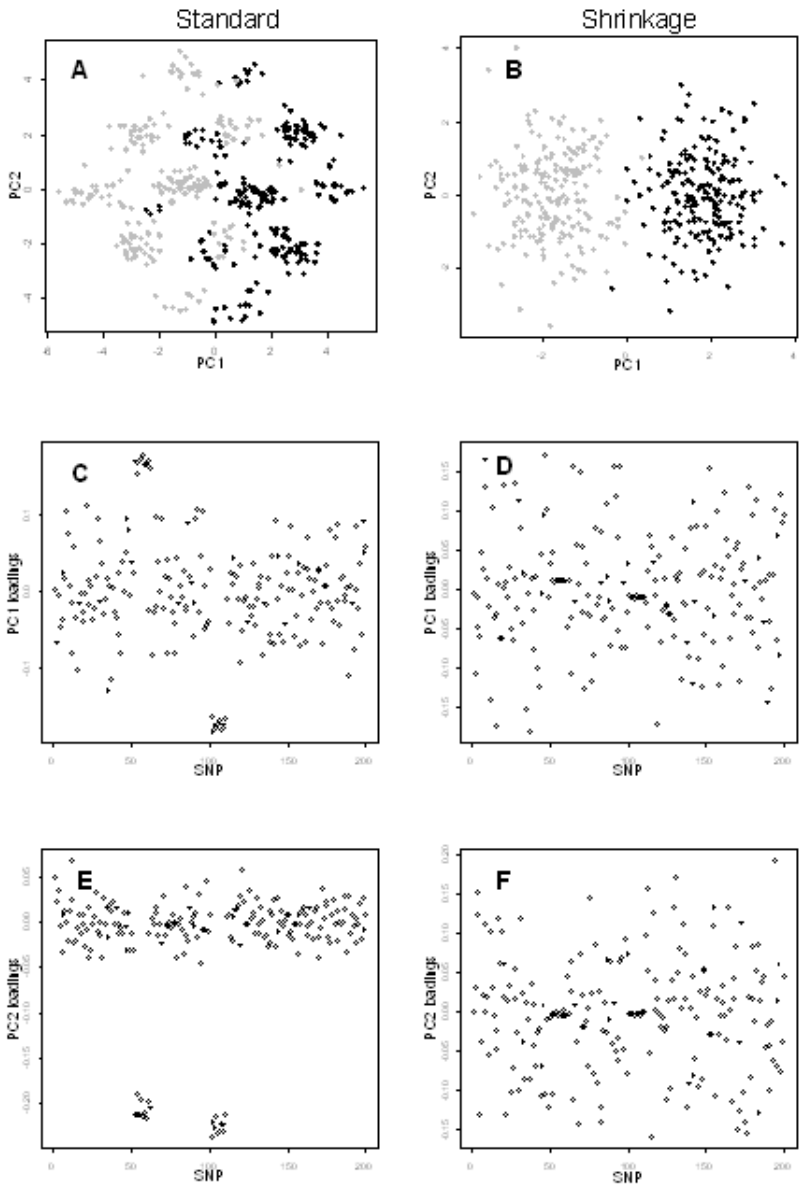

Figure 2.2: Simulation 1 (markers in LD) . Standard PCA (left panels) vs. shrinkage PCA (right panels) in analysis of a stratified population with independent SNPs and two groups of highly dependent SNPs. The different subpopulations are indicated in gray and black. Distinct clumps appear in standard PCA (panel A) which might be falsely interpreted as subpopulations. Panels A and B give the scatter plots of PC1 vs PC2 for the two approaches, while C through F display the loadings of PC1 and PC2, respectively. 1 and 2 are indicators of the original group of each subject.

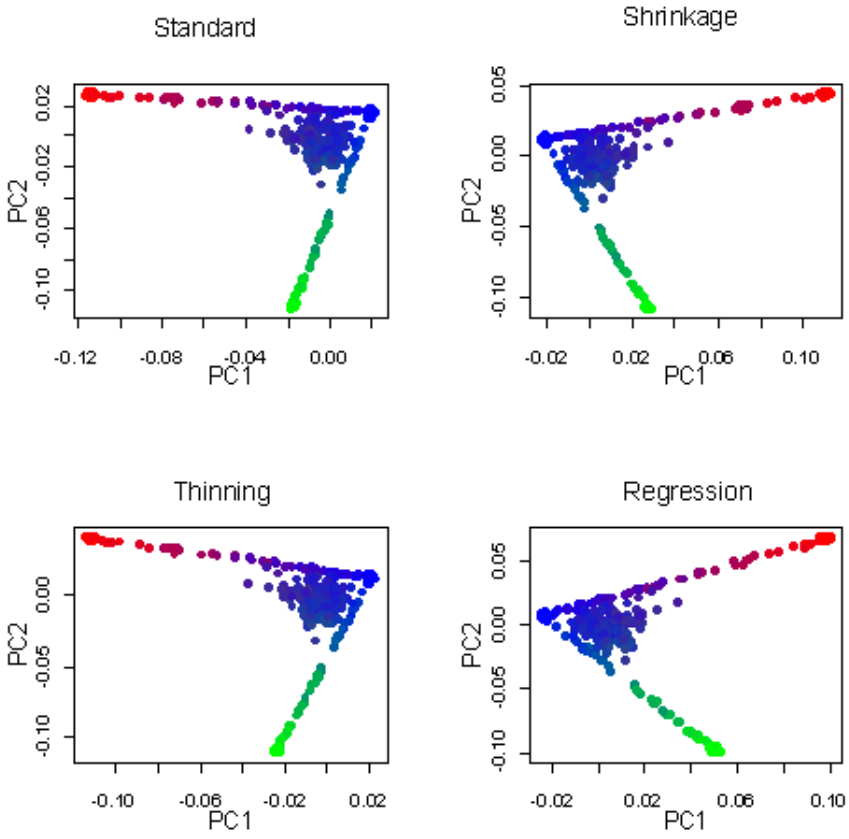

Figure 2.3: Simulation 2 (GWAS data). Scatter plots of the top two PCs of the four PCA methods.

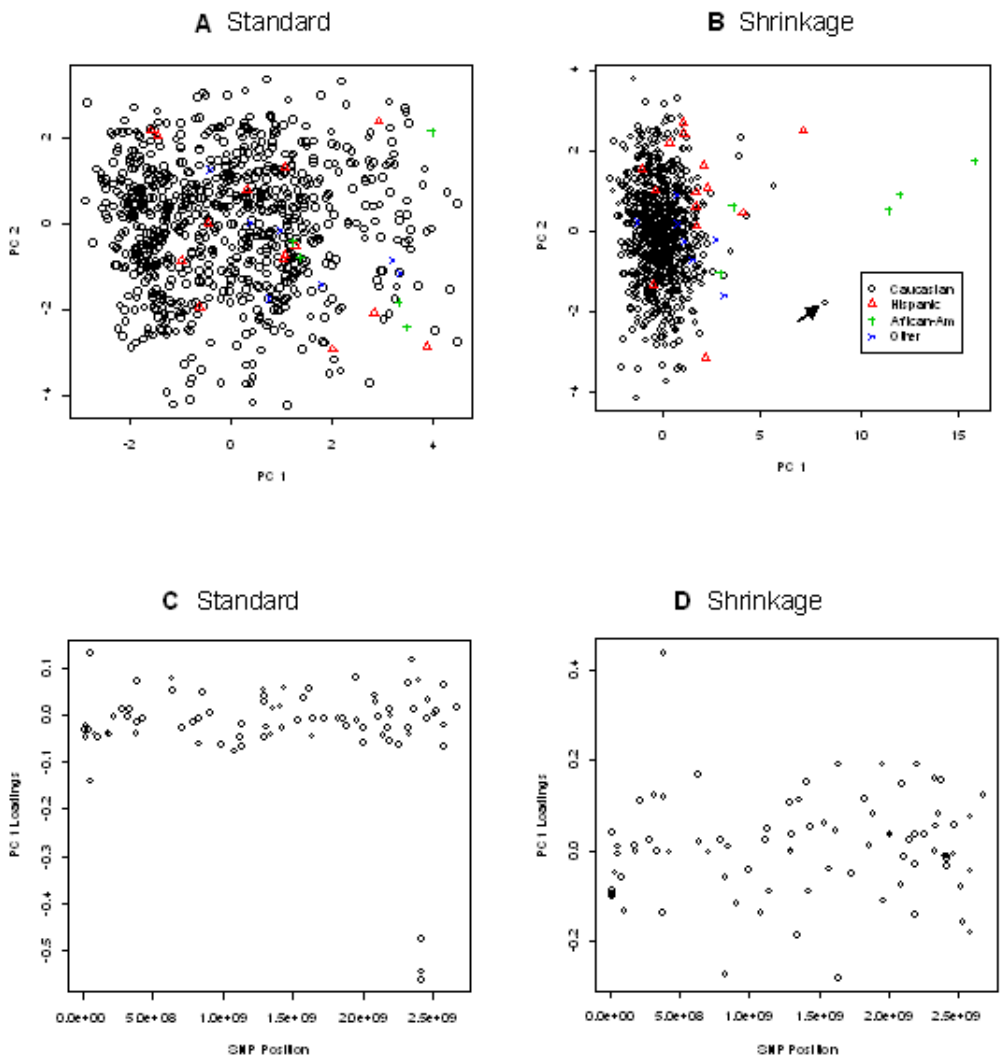

Figure 2.4: Real data analysis 1. Scatter plots of the top two PCs of ancestry-informative markers from the CF Candidate Gene Modifier Study. The left panels are based on standard PCA, while the right panels are from shrinkage PCA.

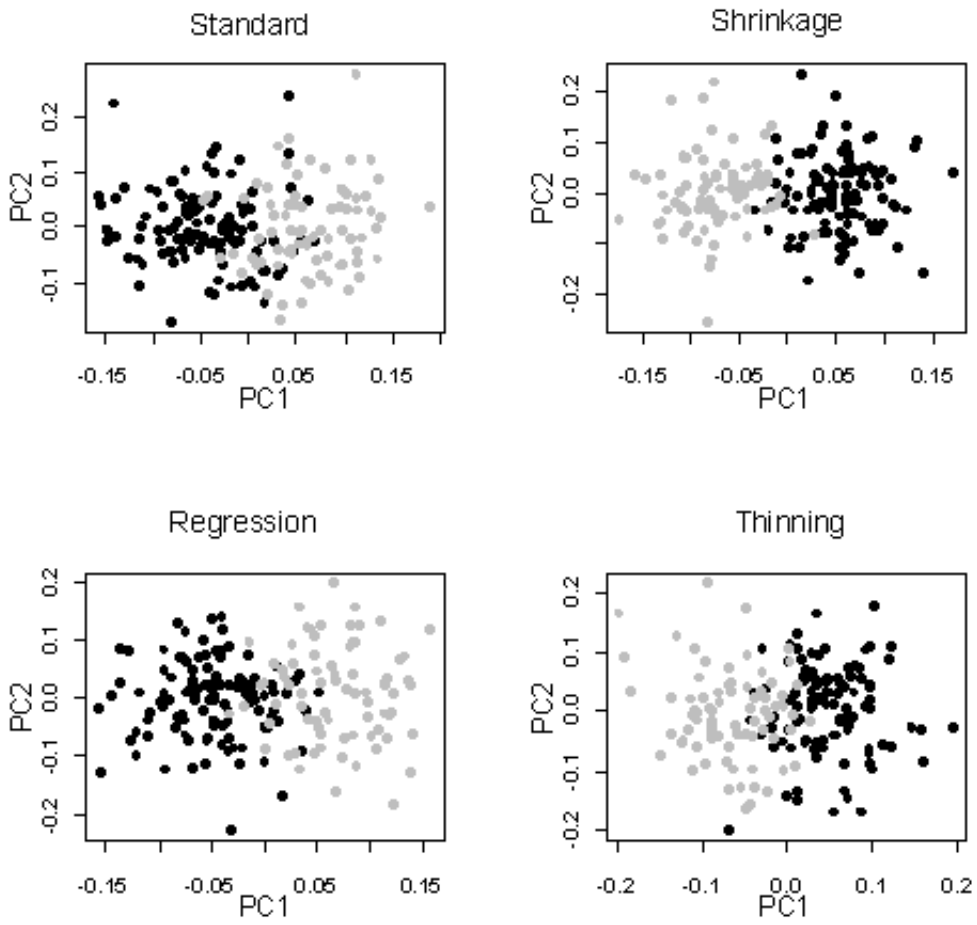

Figure 2.5: Real data analysis 2. Scatter plots of the top two PCs of 4 different methods with different colors for CEU (black) and TSI (grey).

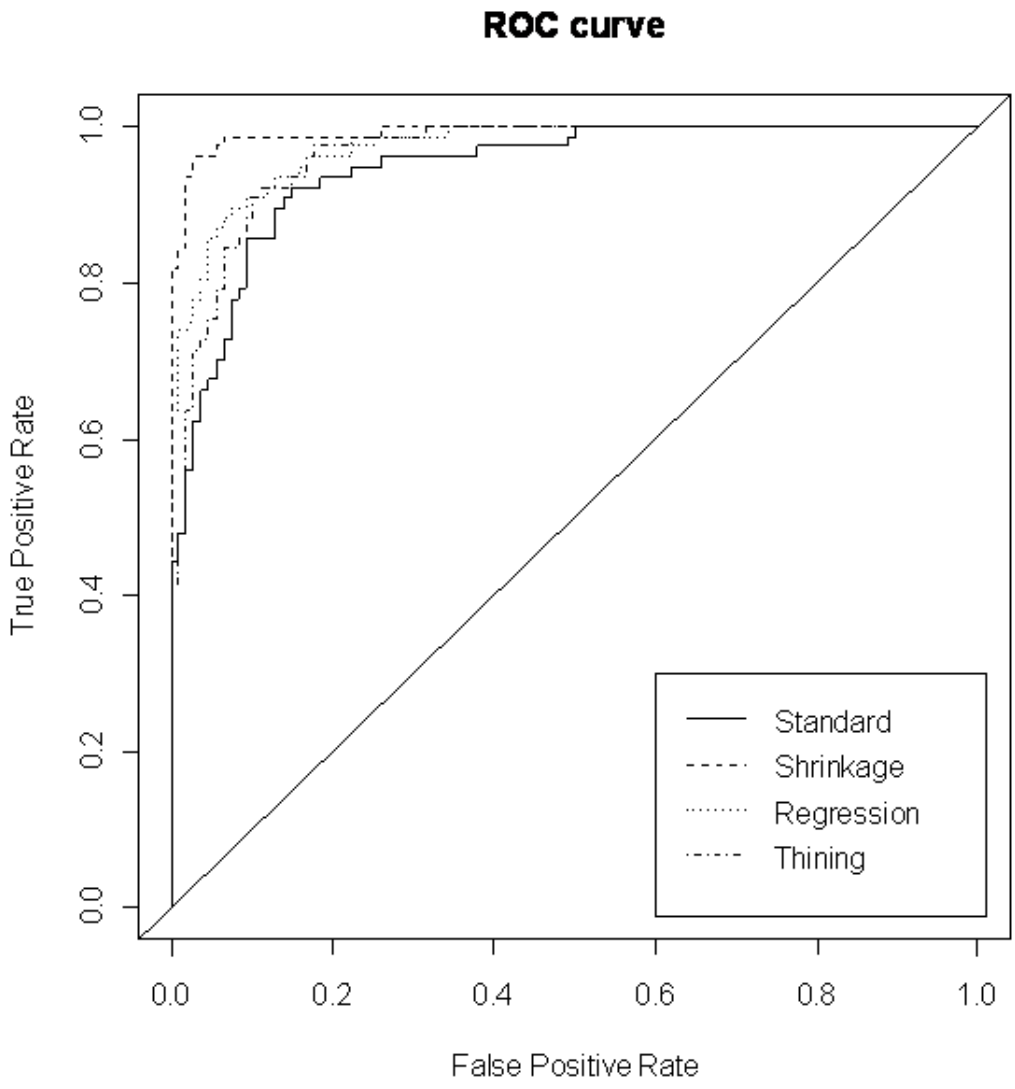

Figure 2.6: Real data analysis 2. ROC curves of using the 1st PC to classify the two sub-populations. 1st PCs were computed from 4 different methods: 1) Standard PCA, 2) Shrinkage PCA, 3) Regression PCA, and 4) Thinning PCA.

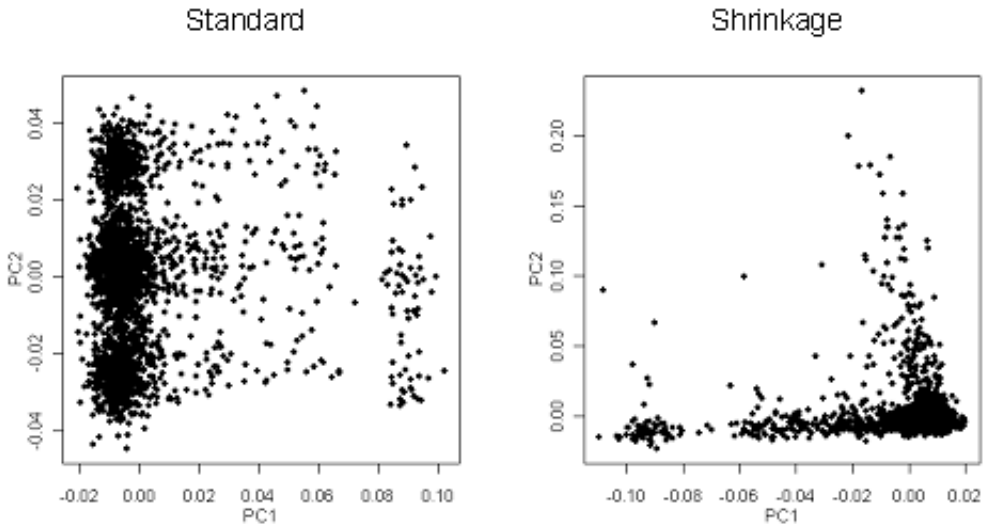

Figure 2.7: Real data analysis 3. Scatter plots of the top two PCs. The left panel is based on standard PCA, while the right panel is from shrinkage PCA.

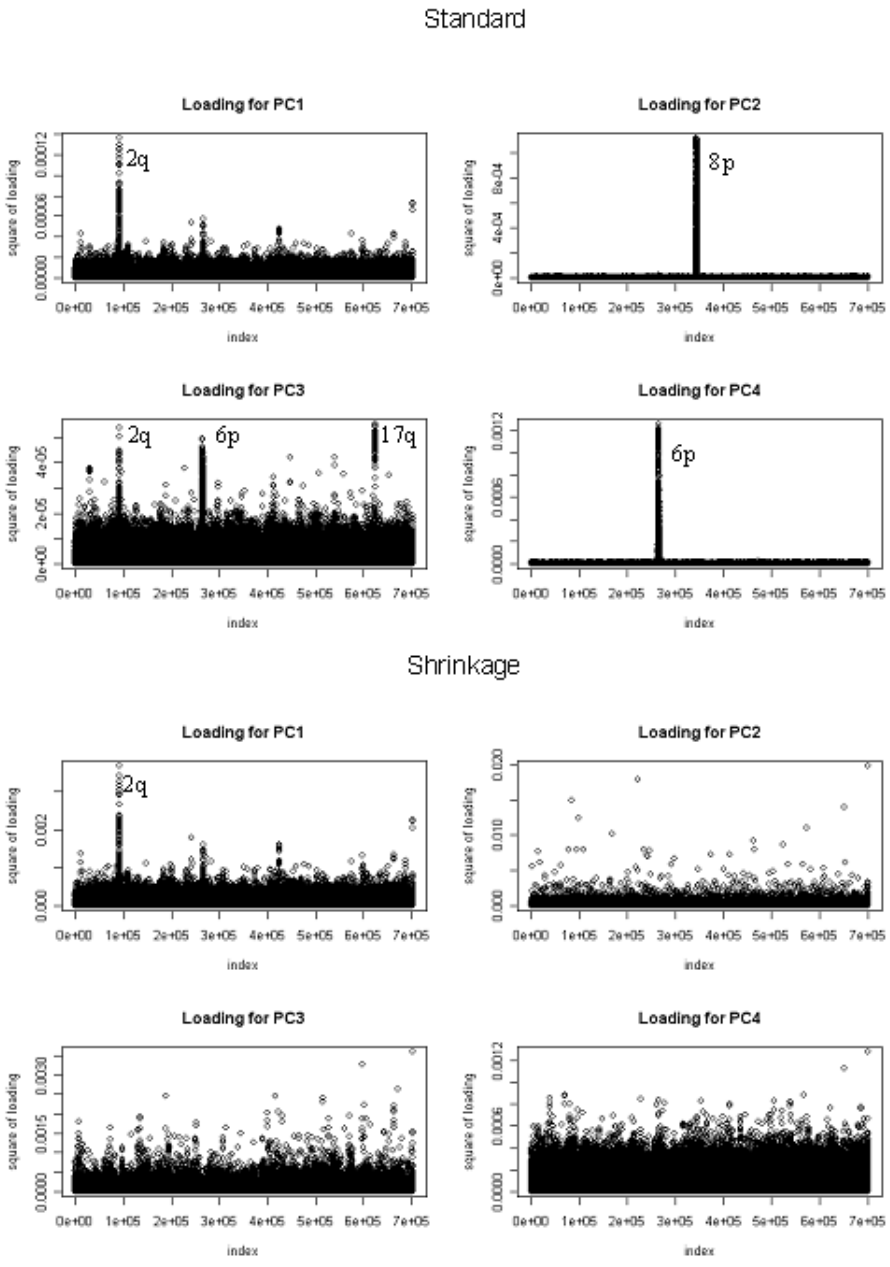

Figure 2.8: Real data analysis 3. Loadings of the top four PCs are displayed for standard PCA (top four panels) and the shrinkage PCA (bottom four panles). The x-axis refers to the serial SNP order on the genome rather than actual physical position.

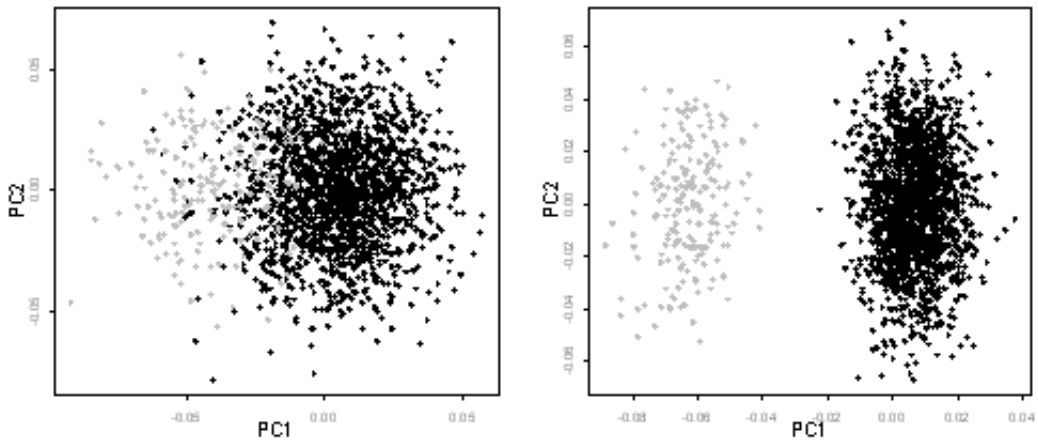

Figure 2.9: Type I and Power issues. Scatter plots of the top two PCs with and without shrinkage. Panel A, standard PCA with no shrinkage. Panel B, shrinkage PCA. The different subpopulations are indicated in gray and black.

Chapter 3

Control of population stratification by correlation-selected principal components

In genome-wide association studies, population stratification is recognized as producing inflated type I error due to the inflation of test statistics. Principal component-based methods applied to genotypes provide information about population structure, and have been widely used to control for stratification. Here we explore the precise relationship between genotype principal components and inflation of association test statistics, thereby drawing a connection between principal component-based stratification control and the alternative approach of genomic control. Our results provide an inherent justification for the use of principal components, but call into question the popular practice of selecting principal components based on significance of eigenvalues alone. We propose a new approach, called EigenCorr, which selects principal components based on both their eigenvalues and their correlation with the (disease) phenotype. Our approach tends to select fewer principal components for stratification control than does testing of eigenvalues alone, providing substantial computational savings and improvements in

power. Analyses of simulated and real data demonstrate the usefulness of the proposed approach.

3.1 Introduction

In tests of genetic association among unrelated individuals, it is recognized that population stratification can result in test statistics with inflated apparent significance, resulting in overall type I error that is far above the nominal level. The method of *genomic control* (Devlin and Roeder, 1999; Devlin, Roeder and Wasserman, 2001; Bacanu, Devlin and Roeder, 2002) was among the first attempts to address this problem. The principle of genomic control is very straightforward. For (chisquare) statistics at numerous markers measuring association with phenotype, an estimate is obtained for the inflation of test statistics beyond that expected under the null hypothesis and assuming no stratification. Then the test statistics are all adjusted by the inflation factor. However, a typical genome-wide association scan (GWAS) tests a very large number of SNP markers, requiring a stringent genome-wide significance threshold. In this setting, genomic control can fail to properly control the type I error (Devlin, Bacanu and Roeder, 2004; Marchini et al., 2004; Zhang, Wang and Deng, 2008), in part because of violations of the assumption of constant variance inflation across the SNPs.

Alternatively, the principal component (PC) approach (Price et al., 2006) uses PCs computed from all genotypes as covariates in phenotype-genotype regression or in stratified analyses. Although the PC approach can be applied without explicit examination of the underlying population substructure, results from numerous studies indicate that the PC values often reflect known substructure and ancestry (Price et al., 2008; Tian et al., 2008). One great advantage of the PC approach is its potential ability to detect subtle population stratification, and to effectively adjust test statistics for only those

markers contributing to the stratification. However, several challenges remain for effective PC-based stratification control. The primary challenge lies in choosing which PCs to include as covariates. Clearly not all PCs can be included, as there are as many PCs as there are individuals under study. Price et al. (2006) originally suggested to use the 10 PCs with the highest eigenvalues. These investigators later proposed using the Tracy-Widom (TW) statistic (Patterson, Price and Reich, 2006) to assess statistical significance of eigenvalues in order to select PCs. However, this approach may detect a very large number of PCs as significant, with uncertain impact on the association analysis. Moreover, the precise contribution of each PC to the overall type I error has not been established. As we shall see below, it is entirely possible for a relatively low-ranked PC to have a greater impact on type I error than does a higher-ranked PC.

The paper is arranged as follows. In Section 3.2, we establish a relationship between PCs and the average of the test statistics. Based on this relationship, we propose a new method, EigenCorr, for selecting PCs based on their corresponding eigenvalues as well as their correlations with the phenotype of interest. The explicit use of phenotypes in stratification control has been anticipated in previous work, e.g. Epstein, Allen and Satten (2007) and Kimmel et al. (2007). However, EigenCorr provides a more direct connection to the type I error inflation introduced by PCs than other approaches of which we are aware. A straightforward generalization of EigenCorr applies to situations where only a subset of markers are used for stratification control. In Section 3.3, we demonstrate the usefulness of EigenCorr via simulation and real GWAS analysis. In Section 3.4, we conclude with a discussion of implications and future directions.

3.2 Materials and Methods

Let g_{ij} be the genotype of SNP i and individual j, where $i = 1, ..., M$ and $j = 1, ...N$. We define a normalized genotype x_{ij} as

$$x_{ij} = \frac{g_{ij} - \overline{g}_{i.}}{\sqrt{\sum_{j=1}^{N}(g_{ij} - \overline{g}_{i.})^2/N}},$$

where $\overline{g}_{i.} = \sum_{j=1}^{N} g_{ij}/N$. Let $\mathbf{X}$ be the resulting $M \times N$ normalized genotype matrix, and $\mathbf{x}_{i.}$ the ith row of $\mathbf{X}$. We have $\sum_{j=1}^{N} x_{ij} = 0$ and $\sum_{j=1}^{N} x_{ij}^2 = 1$. For mathematical precision in later development, the normalization used here is slightly different from that used in Price et al. (2006). However, PCs derived from the two normalizations are nearly identical. From the singular value decomposition (SVD) we obtain $\mathbf{X} = \mathbf{UDP}^T$, where $\mathbf{D}$ is an $N \times N$ diagonal matrix of ordered singular values with jth diagonal element d_j, $\mathbf{U}$ is an $M \times N$ loading matrix, and $\mathbf{P}$ is the $N \times N$ normalized principal component matrix. Let $\boldsymbol{p}_{.j}$ be the jth column of $\mathbf{P}$, (*i.e.*,the jth PC), for $j \in \{1, ..., N\}$. Note that $\boldsymbol{p}_{.j}^T\boldsymbol{p}_{.j} = 1$, $\boldsymbol{p}_{.j}^T\boldsymbol{p}_{.k} = 0$ for $k \neq j$, and $\boldsymbol{p}_{.j}^T\mathbf{1} = 0$ for $j \in \{1, .., N-1\}$ where $\mathbf{1} = \{1, .., 1\}$. Finally, we use $\boldsymbol{y}$ to denote the vector of N phenotypes.

Theorem 1 *Let* $\gamma_j = \boldsymbol{p}_{.j}^T\boldsymbol{y}$. *We have*

$$\sum_{i=1}^{M}(\boldsymbol{x}_{i.}^T\boldsymbol{y})^2 = \sum_{j=1}^{N}\gamma_j^2\lambda_j,$$

where $\lambda_j = d_j^2$ *is the* jth *eigenvalue of* $\mathbf{X}^T\mathbf{X}$.

The proof is given in the Section 3.5.1. As γ_j is an inner product between the (normalized) $\boldsymbol{p}_{\cdot j}$ and $\boldsymbol{y}$, it is easy to show that

$$\gamma_j = \sqrt{\sum_{k=1}^{N}(y_k - \overline{y})^2} \times \text{corr}(\boldsymbol{p}_{\cdot j}, \boldsymbol{y}), \tag{3.1}$$

where "corr" is the Pearson correlation coefficient. Thus γ_j is proportional to the correlation between the phenotype $\boldsymbol{y}$ and the jth PC. We emphasize that the correlation is a *sample* quantity which is observable from the data. Similarly, each term $\boldsymbol{x}_{i\cdot}^T\boldsymbol{y}$ in the equality is proportional to the correlation between the genotype at SNP i and the phenotype. The importance of the result lies in the explicit connection between these M genotype-phenotype correlations to the N PC-phenotype correlations.

3.2.1 Relationship between Genomic Control and Principal Components

Here we obtain explicit results for the relevant test statistics applied in genetic association mapping. An exact correspondence to Theorem 1 technically applies to score test statistics. However, as we demonstrate further below, the results also apply to other common choices of test statistic.

Quantitative Traits

For continuous quantitative phenotype Y, we assume a simple linear regression model at each SNP i:

$$y_j = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_j, \epsilon_j \sim N(0, \sigma^2). \tag{3.2}$$

To test an association of the SNP i with the phenotype, we can use the following score

test (Lehmann and Romano, 2005):

$$S_i = \frac{(\boldsymbol{x}_{i\cdot}^T \boldsymbol{y})^2}{\sum_{j=1}^{N} (y_j - \overline{y})^2 / N}. \tag{3.3}$$

Under the assumption of no genetic effect and no population stratification, S_i is asymptotically distributed as χ^2 with 1 degree of freedom, which has a mean 1. By Theorem 1,

$$\frac{1}{M}\sum_{i=1}^{M} S_i = \left(M \sum_{j=1}^{N} (y_j - \overline{y})^2/N\right)^{-1} \sum_{j=1}^{N} \gamma_j^2 \lambda_j = \frac{N}{M} \sum_{j=1}^{N} \mathrm{corr}^2(\boldsymbol{p}_{\cdot j}, \boldsymbol{y})\ \lambda_j. \tag{3.4}$$

The observed mean of all score test statistics across the M SNPs is then proportional to the sum of the squared PC-phenotype correlations multiplied by their respective eigenvalues.

In a justification of genomic control, it has been argued that under certain models of population stratification, the inflation of test statistics should be similar across all "null" SNPs (Devlin, Roeder and Wasserman, 2001). For this and related work (Devlin and Roeder, 1999), the authors suggested comparing the sample median of test statistics to the chi-square median value, for an estimated inflation factor $\hat{\tau} = \mathrm{median}(S)/0.456$. This approach is intended to be robust to outlying test statistics which presumably correspond to "alternative" SNPs. Other work (Reich and Goldstein, 2001; Devlin, Bacanu and Roeder, 2004) suggests using the sample mean $\hat{\tau} = \bar{S} = (1/M)\sum_{i=1}^{M} S_i$ directly. In practice, the use of the median or mean typically gives similar results, as the proportion of alternative SNPs is typically small. Using either approach, for each i a new statistic $S_i' = S_i/\hat{\tau}$ is then compared to χ_1^2.

To summarize, the results above provide a direct relationship between the mean version of the genomic control quantity $\hat{\tau}$ (left-hand side of (3.4)) and the PC-phenotype

correlations and eigenvalues. This relationship is more than a simple curiosity. While it is known that distinct subpopulations can be represented using PCs (Price et al., 2008), we are not aware that a natural relationship has been previously established between the PCs and the testing procedures. Moreover, (3.4) is exact, holding for any $\mathbf{X}$ and $\boldsymbol{y}$. In particular, this implies that the relationship holds regardless of the underyling population substructure and the proportion of null vs. alternative SNPs. Another point, perhaps more subtle, is that the result is not an expectation, but holds for any realized dataset. Thus the right-hand side of (3.4) is subject to the same sampling variation as $\bar{S}$. Finally, we note that, to the extent that increases in $\bar{S}$ above 1 determine overall inflation of type I error, the equation specifically highlights the terms $\text{corr}^2(\boldsymbol{p}_{\cdot j}, \boldsymbol{y})\ \lambda_j$ as contributors to this inflation. For a principal component to contribute meaningfully to this inflation, it must have both an appreciable λ_j *and* a reasonably large squared sample correlation $\text{corr}^2(\boldsymbol{p}_{\cdot j}, \boldsymbol{y})$. Due to sampling variation, all PCs will have observed $\text{corr}^2(\boldsymbol{p}_{\cdot j}, \boldsymbol{y}) > 0$ for each j, even if the PCs are truly uncorrelated with the population from which $\boldsymbol{y}$ is drawn. The eigenvalues $\lambda_1, ..., \lambda_{N-1}$ are also non-zero. Thus we must distinguish among terms according to their magnitude, and considering sampling variation. Our general approach in the later sections will be to (i) re-rank the PCs by the terms $\text{corr}^2(\boldsymbol{p}_{\cdot j}, \boldsymbol{y})\ \lambda_j$, (ii) test for the statistical significance of each of the terms, and (iii) control for stratification using only those PCs with significant terms.

Case-Control Trait:

Before proceeding further, we establish that the relationships described above are also applicable to case-control studies, with $Y = 0$ and $Y = 1$ corresponding to control and case status, respectively. The data can be analyzed using the logistic regression model (Agresti, 2002) for each SNP i:

$$\log\big(P(Y = 1)/(1 - P(Y = 1))\big) = \beta_{0i} + \beta_{1i} x_{ij}, \tag{3.5}$$

which is conditional on the sampling scheme. Denoting the number of cases as N_1 and the number of controls as N_0, the score test statistic (Agresti, 2002) $S_i = (\boldsymbol{x}_{i.}^T \boldsymbol{y})^2/((N_0N_1)/N^2)$ may be used. However, it is simple to show that $((N_0N_1)/N^2) = \sum_{j=1}^{N}(y_j - \bar{y})^2/N$, and so by comparison with (3.3), we see that (3.4) directly applies.

3.2.2 The influence of stratification on the test statistic at a single SNP

In genomic control, the inflation of test statistics is effectively assumed to be constant across all null SNPs. However, the inflation effect of PCs on the test statistics can be investigated at the level of each SNP. The subsection is intended to be conceptual– in practice we employ the PC-based procedure rather than attemping marker-specific genomic control. For simplicity, we first assume that the individuals are sampled from two subpopulations and the PC analysis fully recovers the two subpopulations via $\boldsymbol{p}_{.1}$. That is, the "true" null model under population stratification is $y_j = \eta_0 + \eta_1 p_{j1} + \epsilon_j$, with $\epsilon_j \sim N(0, \sigma^2)$. Here η_1 refers to the subpopulation effect on the phenotype. The test statistic S_i at the ith SNP does not acknowledge the stratification, and is approximately distributed as

$$\frac{\sigma^2}{\eta_1^2/N + \sigma^2}\ \chi_1^2\big(\frac{u_{i1}^2\eta_1^2\lambda_1}{2\sigma^2}\big) \tag{3.6}$$

where $\chi_1^2(\delta)$ is the non-central chi-square distribution with a noncentrality parameter δ and 1 degree of freedom, and u_{ij} is the $(i,j)^{th}$ element of the loading matrix $\mathbf{U}$ (see Section 3.5.2 for details). The expected value (i.e., the inflation) of S_i can be shown from this result to be $(\sigma^2 + u_{i1}^2\eta_1^2\lambda_1)/(\sigma^2 + \eta_1^2/N)$.

If $\boldsymbol{p}_{.1}$ is included as a covariate in the analysis via the following model at each SNP i,

$$y_j = \eta_0 + \eta_1 p_{j1} + \beta_i x_{ij} + \epsilon_j,$$

we should not expect inflation of the statistic for testing $H_0 : \beta_i = 0$. If there exist $K + 1$ subpopulations in the data, which can be inferred from $\boldsymbol{p}_{.1}, \cdots, \boldsymbol{p}_{.K}$, then the "true" null model can be expressed as

$$y_j = \eta_0 + \sum_{k=1}^{K} \eta_k p_{jk} + \epsilon_j,$$

and the inflation factor of S_i is

$$\frac{\sigma^2 + \sum_{k=1}^{K} u_{ik}^2 \eta_k^2 \lambda_k}{\sigma^2 + \sum_{k=1}^{K} \eta_k^2 / N},$$

which again is locus-specific. Similar conclusions are also applied to dichotomous trait models, but the derivations are more complicated (see Section 3.5.3 for details).

3.2.3 EigenCorr : An Eigenvalue and Correlation-Based PC Selection Procedure

Using the result that the effect of $\boldsymbol{p}_{.j}$ on the mean test statistic is proportional to $\gamma_j^2 \lambda_j$, we propose to select the most significant PCs associated with the population stratification based on the $\gamma_j^2 \lambda_j$, which we call the *EigenCorr* scores. To determine the significance of a given PC, we describe two procedures, which differ in their assumptions concerning population stratification.

1) Method 1: EigenCorr1: We adopt the null hypothesis that the population correlation of the PCs and phenotypes is zero. Furthermore, we assume that there is no population substructure. Under these assumptions, we are able to directly estimate the null distribution of the EigenCorr scores according to the Tracy-Widom distributional approximation (Johnstone, 2001) and the Fisher z-transformation applied to sample

correlations (Fisher, 1921). Specifically, after appropriate normalization the largest eigenvalue λ_1 approximately follows the Tracy-Widom distribution (Patterson, Price and Reich, 2006). We compute $(L_1 - \mu)/\xi$, where

$$L_1 = \frac{(N-1)\lambda_1}{\sum_{k=1}^{N-1} \lambda_k}, \qquad \mu = \frac{\left(\sqrt{m'-1} + \sqrt{N-1}\right)^2}{m'},$$

$$\xi = \frac{\left(\sqrt{m'-1} + \sqrt{N-1}\right)}{m'} \left(\frac{1}{\sqrt{m'-1}} + \frac{1}{\sqrt{N-1}} \right)^{1/3}, \text{ and}$$

$$m' = \frac{N \left(\sum_{k=1}^{N-1} \lambda_k \right)^2}{\left((N-2) \sum_{k=1}^{N-1} \lambda_k^2 \right) - \left(\sum_{k=1}^{N-1} \lambda_k \right)^2}.$$

For other eigenvalues, similar procedures can be followed (see (Patterson, Price and Reich, 2006) for details).

The Fisher z-transformation (Fisher, 1921) provides a highly accurate approximation to the distribution of a correlation coefficient, with $z^* = \frac{1}{2}\log(\frac{1+r_j}{1-r_j})$ approximately normal with mean 0 and variance $\frac{1}{N-3}$. Here r_j is the correlation between the phenotype and $\boldsymbol{p}_{.j}$. Therefore, γ_j, which is proportional to r_j, approximately follows the distribution of

$$\sqrt{\sum_{j=1}^{N} (y_j - \overline{y})^2} \; \frac{e^{2Z} - 1}{e^{2Z} + 1}, \tag{3.7}$$

where $Z \sim N(0, 1/(N-3))$. Using these approximations to the distributions of (independent) γ_j^2 and λ_j, we obtain null distributions for each $\gamma_j^2 \lambda_j$ by simulation, which are then used to compute a p-value for each EigenCorr score. The process proceeds sequentially as follows. We simulate a random variable λ_1^* from the distribution of $(\xi T + \mu)\frac{\sum_{k=1}^{N-1} \lambda_k}{N-1}$, where T is a Tracy-Widom random variable, and simulate γ_1^* using the Fisher z-transformation. These provide the null distribution of $\gamma_1^2 \lambda_1$ from which we

obtain the p-value. After excluding the first eigenvalue, we set $N = N - 1$, recompute μ and ξ, and follow the same procedure sequentially to obtain p-values for each of the remaining EigenCorr scores. We use these p-values to select a number of significant PC covariates, acknowledging multiple comparisons by using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR) to a specified value.

2) Method 2: EigenCorr2: In EigenCorr1, we made the assumption of no population stratification. Although this assumption underlies the current Tracy-Widom testing regimen (Patterson, Price and Reich, 2006), we can relax the assumption, recognizing that significant eigenvalues alone are not sufficient to produce inflation of type I error. For EigenCorr2, we assume only that the PCs are uncorrelated with the population phenotype distribution. In order to perform testing, we treat the λ_j values as fixed, and compute the p-values for high values of γ_j^2, obtainable from Equation (3.7).

Although EigenCorr2 is simpler than EigenCorr1, and is shown to perform well in later simulations, both approaches may have value in different situations. In particular, EigenCorr1 tests both eigenvalues and the PC-phenotpe correlations, and may have an advantage in situations where few eigenvalues are truly signficant. For either approach, our experience indicates that a relatively small number of PCs will be chosen for stratification control, which is desirable for both computational and statistical simplicity.

3.2.4 SNP thinning and weighted PC analysis

In the current practice of PC-based stratification control, investigators are often concerned that the inclusion of SNPs in high linkage disequilibrium can produce misleading results. Thus many investigators choose to "thin" out SNPs so that only a subset with lower correlations is used to generate the PCs (e.g (Fellay et al., 2007)). It is easy

to derive a more general approach as follows. Each SNP is given a weight w_i, such that groups of SNPs in high LD are given lower weight. Specific choices of weights are described elsewhere, but we note that the special case of SNP thinning corresponds to $w_i = 0$ (SNP removed) and $w_i = 1$ (SNP retained). Analysis proceeds by creating a weighted genotype matrix $\widetilde{\mathbf{X}} = \mathbf{WX}$, where $\mathbf{W}$ is a diagonal $M \times M$ matrix with ith diagonal element w_i. Then ordinary PC analysis proceeds using $\widetilde{\mathbf{X}}$ instead of $\mathbf{X}$. Let $\widetilde{\boldsymbol{p}}_{\cdot j}$ be the jth principal component based on $\widetilde{\mathbf{X}}$, and $\tilde{\lambda}_j$ its corresponding eigenvalue. By Theorem 1,

$$\sum_{i=1}^{M} (\tilde{\boldsymbol{x}}_{i\cdot}^T \boldsymbol{y})^2 = \sum_{j=1}^{N} \tilde{\gamma}_j{}^2 \tilde{\lambda}_j{}^2$$

where $\tilde{\gamma}_j = \widetilde{\boldsymbol{p}}_{\cdot j}^T \boldsymbol{y}$. Since $\sum_{i=1}^{M} (\tilde{\boldsymbol{x}}_{i\cdot}^T \boldsymbol{y})^2 = \sum_{i=1}^{M} w_i^2 (\boldsymbol{x}_{i\cdot}^T \boldsymbol{y})^2$, the weighted mean of score test statistics is

$$\sum_{i=1}^{M} w_i^2 S_i = \frac{1}{\sum_{j=1}^{N} (y_j - \overline{y})^2 / N} \sum_{j=1}^{N} \tilde{\gamma}_j{}^2 \tilde{\lambda}_j{}^2 \tag{3.8}$$

Here again the EigenCorr procedure can be applied, but to the EigenCorr scores based on the weighted PCs.

In the special case of SNP thinning (weights of 0 or 1), the exact connection to genomic control remains, provided that the genomic control $\hat{\tau}$ is obtained using only the SNPs used for PC analysis. In practice, genomic control is usually performed using all SNPs, while the PCs are calculated using a thinned set of SNPs. However, we demonstrate in later simulations that the approximate relationship still holds.

3.3 Simulations and Real Data Analysis

We investigated the performance of the proposed EigenCorr approach in applications to simulated data and a real GWAS data set.

3.3.1 Simulation Studies

Simulation 1: We simulated 1000 samples from 5 subpopulations with 20,000 uncorrelated SNPs, with 210 samples from each of the first four subpopulations, and the remaining 160 samples from subpopulation 5. For each SNP, the overall minor allele frequency (MAF) was uniform from 0.05 to 0.5, and F_{st} was uniform from 0.01 to 0.04. From these values, the MAF for each subpopulation was generated according to the Balding-Nichols model (Balding and Nichols, 1995*b*). PC analysis showed that the top 4 PCs were significant according to the TW test, with $\boldsymbol{p}_{.4}$ specifically distinguishing subpopulation 5 from the others.

To simulate population stratification, we generated a disease phenotype from logistic regression with log(odds ratio)=1.6 between the samples from subpopulation 5 and the remaining samples. Therefore, increases in type I error resulting from population stratification arise entirely from the differing disease prevalence between subpopulation 5 and the remaining samples. Figure 3.1 shows eigenvalues and EigenCorr scores of the first 10 PCs. The TW test selected the top 4 PCs as significant at $p < 0.01$, since its selection is entirely eigenvalue based, while only $\boldsymbol{p}_{.4}$ was identified, correctly, by EigenCorr. This simulation is illustrative of the intended advantage of the use of EigenCorr scores.

Simulation 2: simulations based on a real dataset. To investigate type I error and power associated with PC-based methods, we simulated phenotypes based on a real schizophrenia GWAS study from the GAIN consortium [Version 2, Accession number: phs000021.v2.p1] (Sanders et al., 2008). In this manner we intended to reflect the genetic complexity encountered in real studies. The filtered General Research Use (GRU) African American data consists of 1904 samples with 845,814 SNPs, and was downloaded from dbGap at NCBI (ncbi.nlm.nih.gov). We first filtered out highly related or duplicated samples, obtaining 1847 samples. Markers with a high rate of

missing values (> 0.05), or a low minor allele frequency (< 0.01), were excluded from the analysis. In addition, markers on sex chromosomes were excluded, resulting in a final total of 810,264 markers.

To alleviate the effects of linkage disequilibrium among markers, we thinned out SNPs based on pairwise correlation, such that no pair of SNPs had $r^2 > 0.1$. After such SNP thinning, 96,346 SNPs remained. We further removed outliers based on computed PCs. Using TW statistics, we found that 98 PCs had p-values smaller than 0.01. Among these, the first two PCs exhibited a clear clustering indicative of the stratification structure, and we did not use these two PCs to identify outliers. For the remaining 96 PCs, we applied the standard $6\times$ SD rule to identify outliers (Luca et al., 2008), and total 12 subjects were removed. After recalculation, 91 PCs had TW p-values smaller than 0.01.

We used $\boldsymbol{p}_{\cdot 1}, \boldsymbol{p}_{\cdot 2}, \boldsymbol{p}_{\cdot 5}$, and $\boldsymbol{p}_{\cdot 10}$ to generate association of strata with a simulated phenotype. For quantitative traits, under the null hypothesis that no SNP is associated with the phenotype, we simulated phenotypes according to

$$y_j = \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10} + \epsilon_j.$$

Under the alternative hypothesis where there is a disease SNP associated with the phenotype, we used

$$y_j = \beta x + \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10} + \epsilon_j$$

for SNP genotype x and $\beta = 0.15$. For both models, the ϵ_j were generated from a normal distribution with mean 0 and variance 1, $\eta_1, \eta_2, \eta_5, \eta_{10}$ were generated as independent and identically normally distributed so as to contribute the half of the variability of $\boldsymbol{y}$ under the null model, and the same distribution of η was used for the alternative

model. The results were analyzed by linear regression.

Similarly, for a dichotomous phenotype, we used the model

$$\log\big(P(Y_j = 1)/(1 - P(Y_j = 1))\big) = \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10}$$

and

$$\log\big(P(Y_j = 1)/(1 - P(Y_j = 1))\big) = \beta x + \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10}$$

under the null and alternative hypotheses respectively, with $\beta = 0.47$. The coefficients $\eta_1, \eta_2, \eta_5, \eta_{10}$ were randomly generated from the normal distribution to make the variance of log odds ratio 4.0 under the null model, and the same distribution of η was used for the alternative model. The choice of genotype effect β corresponds to a relatively strong allelic odds ratio of 1.6. The results were analyzed by logistic regression and testing by maximum likelihood ratios.

The TW test selected 91 PCs at FDR level 0.1, the same PCs having TW p-values $<$ 0.01. Note that the genotypes were fixed, but the phenotypes, and therefore the results of EigenCorr1 and EigenCorr2, varied across simulations. On average, EigenCorr1 and EigenCorr2 selected 3.52 and 3.51 PCs at FDR level 0.1 among the first 200 PCs, respectively, with most of them overlapping and typically reflecting the true PC stratification.

Table 3.1 provides empirical type I errors for the quantitative and case-control data, separately for the 5 different approaches: 1) no adjustment for population stratification; 2) adjustment by the variables $\boldsymbol{p}_{.1}, \boldsymbol{p}_{.2}, \boldsymbol{p}_{.5}$ and $\boldsymbol{p}_{.10}$ representing the true population stratification; 3) adjustment by the 91 PCs selected by the TW test; 4) adjustment by the PCs selected by EigenCorr1, and 5) adjustment by EigenCorr2. Method 2 can be viewed as a gold standard, where the true population stratification is assumed known

and modeled. P-values were computed using a likelihood ratio test (LRT). The same simulations with values from the score test are given in Table 3.2.

From the tables we can see that adjustment of population stratification is necessary to control the type I error. Adjustment by the known confounding PCs controls for type I error, as expected. Adjustment by the 91 top PCs controls type I error for the quantitative trait, as would be expected in a linear model in which the normal model assumptions hold. However, adjustment by the TW PCs results in somewhat inflated type I error for the dichotomous trait. In practice, investigators might be reluctant to fit so many covariates, but the absence of a principled alternate procedure based solely on eigenvalues makes it difficult to prescribe an alternative, when so many eigenvalues are clearly significant. In contrast, both Eigencorr methods provide proper type I error control.

In terms of statistical power, both Eigencorr methods are comparable to knowing the true confounding PCs, and are more powerful than PC selection based on the TW test.

Table 3.3 describes the estimated genetic effect β for the five methods. All five methods give essentially unbiased estimates of the genetic effect for the quantitative trait. The logistic regression estimates, however, are biased below the true value under no adjustment, and upwardly biased when the TW PCs are used. The presence of both bias and poorly-controlled type I error is a well-known consequence of the inclusion of unnecessary covariates in logistic regression (Lubin, 1981). Finally, we mention that simulations for the setups in Tables 3.1 and 3.3 were also performed for the situation where (i) no SNPs were thinned for PC analysis, (ii) no outlier detection was performed, and (iii) a simple testing criterion for EigenCorr was performed, including any Eigen-Corr scores/ correlations with nominal p-values < 0.01. The results (not shown) were all qualitatively similar to the results described here. Although it is not possible to

cover every scenario, we believe the results are relatively robust to reasonable choices for data pre-treatment and PC covariate testing.

Simulation 3: stringent testing thresholds. Genome scans typically use very stringent testing thresholds in order to control the overall family-wise error rate. To investigate type I errors under such stringent thresholds, we conducted additional simulations of 10^9 p-values for each setup described above for simulation 2 on the schizophrenia dataset, using thinned PCs and for which outliers had been removed. To do so, we largely followed the null simulation procedure described for simulation 2. However, in order to make the computation feasible, we divided the genotypes into 16 portions, each with approximately 50,000 SNPs. For each portion, we randomly generated $\boldsymbol{y}$ under the null and performed the regressions as described, and continued until a total of 10^9 p-values were obtained. This approach is unbiased, but retains modest correlation of p-values within the portions. For these simulations we computed results from the score tests only, as this approach does not require estimating parameters under the alternative, and thus faster computation can be performed. The results are given in Table 3.4 for a series of thresholds, down to 10^{-7}. The results show that the large number of PCs from the TW test inflates type I error for the dichotomous trait. For nominal (SNP-specific) p-value thresholds of 10^{-6} and 10^{-7}, which are of the order used in genome scans, the type I error is inflated to twice the nominal level. Note also that for the dichotomous trait, adjustment by the known confounding PCs is somewhat conservative, which occurs due to the extreme threshold and the finite sample size. Using results from the known confounding PCs as a gold standard, both EigenCorr1 and EigenCorr2 show proper type I error, and are slightly conservative when compared to the nominal intended threshold.

3.3.2 The GAIN Schizophrenia Data

We next applied the EigenCorr methods to the schizophrenia data, using the actual dichomotous case-control phenotype and logistic regression. PCs were computed using the same procedures for obtaining thinned PCs and excluding outliers as described for simulation 2. Among the first 200 PCs EigenCorr1 selected 2 PCs at FDR = 0.1, and EigenCorr2 selected 4 PCs at this level, including the 2 PCs selected by EigenCorr1. Importantly, some of the PCs with significant EigenCorr scores had relatively low-ranking eigenvalues, which would not have been captured under simple prescriptions such as using PCs with the top 10 eigenvalues. For the top 91 PCs selected by the TW test, Figure 3.2 displays their eigenvalues, their correlations with the trait, and their EigenCorr scores. Clearly many PCs are essentially uncorrelated with the trait. We tested the association of each genetic marker with the trait using the logistic regression LRT and all methods described in the earlier simulation section, except that the "known confounding PCs" method was not possible. Figure 3.3 shows a QQ plot of the $-\log_{10}$ (p-value) (observed vs. expected) from the 4 methods with the 95% prediction band (Stirling, 1982). The QQ plot with no population stratification adjustment deviates dramatically from the diagonal. The QQ plot using the TW adjustment also shows deviation from the diagonal, likely caused by a large number of unnecessary PCs in the model. In contrast, QQ plots from both EigenCorr1 and EigenCorr2 suggest proper type I error control, and no SNP reached genome-wide significance.

The inclusion of unnecessary covariates can also have a substantial computational cost, involving optimization and matrix inversion for large numbers of parameters. Using the 91 PCs selected by TW, computation of LRT for the whole dataset took 9.1 hours using R (http://www.r-project.org/) on 11 node Linux clusters, where each node operates at 2.6 GHz with 8 gb RAM. In contrast, analysis with the 2-4 selected PCs from EigenCorr took approximately 1 hour.

To illustrate the impact of outliers and LD for each method, we computed PCs without SNP pruning and outlier exclusion, and selected PCs based on TW test and EigenCorr scores. Among the first 200 PCs and using an FDR threshold of 0.1, the TW test selected the first 164 PCs, while EigenCorr1 selected 7 PCs and EigenCorr2 selected 8 PCs, including the 7 PCs selected by EigenCorr1. QQ plots are shown in Figure 3.4. The figure shows that EigenCorr can control type I error while selecting a reasonably small number of PCs, even in the presence of LD and outliers.

3.3.3 An empirical comparison of association statistics, and the impact of SNP thinning

In the main result of this paper, we showed that the mean of the score test statistics is proportional to the sum of the EigenCorr scores. Among of the most fundamental results in statistics is the asymptotic equivalence of score statistics, likelihood ratio statistics, and Wald statistics in a suitable neighborhood of the null (Lehmann and Romano, 2005). To illustrate the applicability of our results to other choices of statistic, we compared the standard score statistics for SNP effect for the GAIN consortium schizophrenia dataset to the likelihood ratio and Wald statistics with no additional covariates. In addition, we performed 10 permutations of the phenotypes relative to the genotypes and recomputed the entire set of statistics across all SNPs for each permutation. Figure 3.5 (panel A) shows that in both the original and permuted datasets the mean score statistic indeed follows the equality in equation (3.3). For this plot, the sum of EigenCorr scores was computed without thinning, to illustrate the exact match to the mean of the score test statistics for all SNPs. In addition, the mean of the other two statistics is also very nearly on the unit line. Note that the mean of the three statistics follow each other across the simulations, even though there is considerable variation across the simulations. Note that the actual data (grey symbols

on plot) shows more extreme summed EigenCorr scores than the permutations, as is expected for data with true stratification. Furthermore, the median of each of the three statistics follows the same pattern (Figure 3.5, panel B), adhering closely to the predicted line with slope 0.456, corresponding the median of a χ^2_1 distribution.

As we have described, PC computation is usually applied after SNP thinning. We compared the mean and median of the test statistics for the thinned SNPs vs. these quantities for all SNPs (Figure 3.5, panels C and D). Interestingly, the mean and median of the test statistics for the thinned SNPs closely track those for all SNPs, even though the thinned set represents only 12% of all SNPs. We believe this result follows from the fact that the thinned SNPs represent the overall PCs well, and thus also the correlations of phenotypes to PCs. We conclude that, in addition to motivation based on asymptotics, our comparison to the genomic control inflation factor is appropriate for the most common choices of test statistics, with sample sizes and numbers of SNPs encountered in practice.

3.4 Discussion

In this paper we have shown that the average inflation of test statistics is determined by genotype PCs according to their eigenvalues and their correlations with the phenotype. We specifically highlight the EigenCorr scores, $\gamma^2\lambda$, as the quantities of interest for PC-based stratification control, and have clarified that PCs that are uncorrelated with the phenotype are of little concern. The explicit connection to genomic control provides insight into the advantages of PC-based stratification control. Moreover, the results provide a natural motivation for the use of PC-based control that does not depend on specific population assumptions. In addition to the statistical advantages of EigenCorr, the reduced computational time achieved by focusing on only the problematic PCs is an important advantage. This will be especially true for GWAS analyses involving

larger number of markers and samples, and in which the data are analyzed for gene-gene interactions. The reduction in computation afforded by EigenCorr may not be substantial for the score test, since it does not require fitting the alternative model.

In our analyses, EigenCorr1 and EigenCorr2 produced similar results. However, we believe that EigenCorr1 is theoretically more desirable, as it uses the EigenCorr score directly as a test statistic. However, EigenCorr1 does assume that the sample eigenvalues follow the TW distribution. The null distribution of the statistic may not be accurate for the largest eigenvalues, which may depart sharply from the null. Nonetheless, the role of the genotype-phenotype correlations γ^2 still provides a hedge against fitting an excessive number of PCs.

EigenCorr is not the first approach in which the phenotype information has been used to identify or construct genotype-based covariates. Epstein, Allen and Satten (2007) described a general stratification score approach, with the use of partial least squares (PLS) of phenotype on a number of markers as one approach to construct such a score. Lee et al. (2008) have pointed out that PLS can result in overfitting and reduce power, because PLS is designed to maximize the fit to phenotypic variation. However, Epstein, Allen and Satten (2007) and the rejoinder (Epstein, Allen and Satten, 2008) provide important clarification that desirable stratification control should explain some of the true variability in phenotype. This fact was also recognized by Kimmel et al. (2007), who described initial correction of phenotype by a small number of PC clusters before proceeding with a permutation approach for signficance testing by genotype.

Zhao, Rebbeck and Mitra (2009) have computed a propensity score using (genetic) covariates to predict genotype at a test locus before inclusion in phenotype-genotype model. This approach, although implemented with relatively few genetic covariates (serving a role analogous to our PCs), potentially avoids the problem of overfitting in

selecting covariates associated with the phenotype. However, it is potentially susceptible to the influence of an excessive number of genetic covariates, which we have shown can be a problem with logistic regression.

Viewed in this light, the EigenCorr approach may be seen as generally similar to that advocated by Epstein, Allen and Satten (2007), using PC-based phenotype adjustment in the form of regression covariates. However, the EigenCorr procedure, with a fixed number of PCs to choose from, is much less flexible than procedures such as PLS, and the FDR testing procedure provides a natural penalty against overfitting. Moreover, in contrast to other procedures, the EigenCorr motivation and approach is explicitly connected to the source of test statistic inflation. The fact that genotype must also be associated with population stratum in order to create confounding is also implicit in EigenCorr, because each informative marker has an influence on, and will be associated with, at least one eigenvector. We feel that EigenCorr offers an efficient filter to identify the confounding variables of greatest influence.

3.5 Proofs

3.5.1 Proof of Theorem 1

Since $\boldsymbol{p}_{.j}, j = 1, .., N$ are the orthonormal basis of R^N space, $\boldsymbol{y}$ can be expressed as the linear combination of these vectors, $\boldsymbol{y} = \sum_{l=1}^{N} a_l \boldsymbol{p}_{.l}$, where a_l is a coefficient of each vector $\boldsymbol{p}_{.l}$. However, $\gamma_l = \boldsymbol{p}_{.l}^T \boldsymbol{y} = a_l \boldsymbol{p}_{.l}^T \boldsymbol{p}_{.l} = a_l$, resulting in $\boldsymbol{y} = \sum_{l=1}^{N} \gamma_l \boldsymbol{p}_{.l}$.

Now since $\boldsymbol{x}_{i.} = \sum_{k=1}^{N} e_k u_{ik} \boldsymbol{p}_{.k}$, we get

$$\boldsymbol{x}_{i.}^T \boldsymbol{y} = \sum_{k=1}^{N} e_k u_{ik} \boldsymbol{p}_{.k}^T \left(\sum_{l=1}^{N} \gamma_l \boldsymbol{p}_{.l} \right) = \sum_{k=1}^{N} e_k u_{ik} \gamma_k \text{ and}$$

$$\begin{aligned}
\sum_{i=1}^{M}(\boldsymbol{x}_{i.}^T\boldsymbol{y})^2 &= \sum_{i=1}^{M}\left(\sum_{k=1}^{N}e_k u_{ik}\gamma_k\right)^2 = \sum_{i=1}^{M}\sum_{k=1}^{N}e_k^2u_{ik}^2\gamma_k^2 + 2\sum_{i=1}^{M}\sum_{k<l}^{N}e_k e_l u_{ik}u_{il}\gamma_k\gamma_l \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}e_k^2u_{ik}^2\gamma_k^2 + 2\sum_{k<l}^{N}e_k e_l\gamma_k\gamma_l\sum_{i=1}^{M}u_{ik}u_{il} \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}e_k^2u_{ik}^2\gamma_k^2 = \sum_{k=1}^{N}e_k^2\gamma_k^2\sum_{i=1}^{M}u_{ik}^2 = \sum_{k=1}^{N}e_k^2\gamma_k^2 \\
&= \sum_{k=1}^{N}\gamma_k^2\lambda_k,
\end{aligned}$$

since the loading matrix $\mathbf{U}$ is orthogonal, that is $\sum_{i=1}^{M}u_{ik}^2 = 1$ for all k and $\sum_{i=1}^{M}u_{ik}u_{il} = 0$ for $k \neq l$.

3.5.2 Quantitative Trait

Let us assume that the true model of $\boldsymbol{y}$ is $y_j = \eta_0 + \eta_1 p_{j1} + \epsilon_j$, where $\epsilon_j \sim N(0,\sigma^2)$ for $j \in \{1,\cdots,N\}$. The score test statistic for SNP i is

$$S_i = \frac{(\boldsymbol{x}_{i.}^T\boldsymbol{y})^2}{\sum_{j=1}^{N}(y_j-\overline{y})^2/N}$$

The numerator of the score test statistic is

$$\frac{1}{\sigma^2}(\boldsymbol{x}_{i.}^T\boldsymbol{y})^2 = \frac{1}{\sigma^2}(\boldsymbol{x}_{i.}^T(\boldsymbol{y}-\eta_0\mathbf{1}))^2 = \frac{1}{\sigma^2}(\boldsymbol{y}-\eta_0\mathbf{1})^T(\boldsymbol{x}_{i.}\boldsymbol{x}_{i.}^T)(\boldsymbol{y}-\eta_0\mathbf{1}) = \frac{1}{\sigma^2}(\eta_1\boldsymbol{p}_{.1}+\boldsymbol{\epsilon})^T(\boldsymbol{x}_{i.}\boldsymbol{x}_{i.}^T)(\eta_1\boldsymbol{p}_{.1}+\boldsymbol{\epsilon}). \tag{3.9}$$

where $\boldsymbol{\epsilon} = \{\epsilon_j\}$. Since $\boldsymbol{x}_{i.}\boldsymbol{x}_{i.}^T$ is an idempotent matrix with rank 1, $\frac{1}{\sigma^2}(\boldsymbol{x}_{i.}^T\boldsymbol{y})^2$ follows as noncentral chi-square distribution with 1 degrees of freedom and the noncentrality

parameter

$$\mu = \frac{(\eta_1 \boldsymbol{p}_{.1})^T (\boldsymbol{x}_{i.} \boldsymbol{x}_{i.}^T)(\eta_1 \boldsymbol{p}_{.1})}{2\sigma^2} = \frac{\eta_1^2 (\boldsymbol{p}_{.1}^T \boldsymbol{x}_{i.})^2}{2\sigma^2}$$
$$= \frac{\eta_1^2 u_{i1}^2 \lambda_1}{2\sigma^2}$$

The denominator of the score test statistic is

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N}(y_j - \overline{y})^2 &= \frac{1}{N}\sum_{j=1}^{N}(\eta_1 p_{j1} + \epsilon_j - \overline{\epsilon})^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}\eta_1^2 p_{j1}^2 + \frac{1}{N}\sum_{j=1}^{N}(\epsilon_j - \overline{\epsilon})^2 + \frac{2}{N}\sum_{j=1}^{N}\eta_1 p_{j1}(\epsilon_j - \overline{\epsilon}) \\
&= \frac{\eta_1^2}{N} + \frac{1}{N}\sum_{j=1}^{N}(\epsilon_j - \overline{\epsilon})^2 + \frac{2}{N}\sum_{j=1}^{N}\eta_1 p_{j1}\epsilon_j \qquad (3.10)
\end{aligned}$$

By the law of large numbers, $\frac{1}{N}\sum_{j=1}^{N}(\epsilon_j - \overline{\epsilon})^2$ converges in probability to σ^2 and $\frac{1}{N}\sum_{j=1}^{N}\eta_1 p_{j1}\epsilon_j$ converges to 0. By (3.9) and (3.10),

$$S_i \sim \frac{\sigma^2}{\eta_1^2/N + \sigma^2}\chi_1^2\left(\frac{\eta_1^2 u_{i1}^2 \lambda_1}{2\sigma^2}\right)$$

3.5.3 Case-Control Trait

For a dichotomous trait, we need the moderate population effect assumption, which is also required for the genomic control. Let us assume that the true model of Y is $\log\big(P(Y=1)/(1-P(Y=1))\big) = \eta_0 + \eta_1 p_{j1}$. The score test statistic S_i of SNP i is

$$S_i = \frac{(\sum_{j=1}^{N} y_j x_{ij})^2}{r(1-r)} \qquad (3.11)$$

where $r = N_1/N$. By the Lindberg Feller CLT, $\sum_{j=1}^{N} y_j x_{ij}$ follows the normal distribution. Let $\alpha = 1/exp(\eta_0)$, then we can approximate the mean of $\sum_{j=1}^{N} y_j x_{ij}$ by the first order talyor expansion.

$$\begin{aligned} E(\sum_{j=1}^{N} y_j x_{ij}) &= \sum_{j=1}^{N} \frac{exp(\eta_0 + \eta_1 p_{j1})}{1 + exp(\eta_0 + \eta_1 p_{j1})} x_{ij} \approx \sum_{j=1}^{N} x_{ij} (\frac{1}{\alpha + 1} + \frac{\alpha}{(\alpha+1)^2} \eta_1 p_{j1}) \\ &= \frac{\alpha}{(\alpha+1)^2} \eta_1 u_{i1} e_1 \end{aligned} \tag{3.12}$$

Since we assumed moderate population effect, approximation by the first order taylor expansion can be hold. Variance of $\sum_{j=1}^{N} y_j x_{ij}$ is

$$\begin{aligned} Var(\sum_{j=1}^{N} y_j x_{ij}) &= \sum_{j=1}^{N} x_{ij}^2 Var(y_j) = \sum_{j=1}^{N} x_{ij}^2 (\frac{exp(\eta_0 + \eta_1 p_{j1})}{1 + exp(\eta_0 + \eta_1 p_{j1})})(\frac{1}{1 + exp(\eta_0 + \eta_1 p_{j1})}) \\ &\approx \sum_{j=1}^{N} x_{ij}^2 (\frac{1}{\alpha+1} + \frac{\alpha}{(\alpha+1)^2} \eta_1 p_{j1})(\frac{\alpha}{\alpha+1} - \frac{\alpha}{(\alpha+1)^2} \eta_1 p_{j1}) \\ &= \frac{\alpha}{(\alpha+1)^2} + \frac{\alpha^2 - \alpha}{(\alpha+1)^3} \eta_1 \sum_{j=1}^{N} x_{ij}^2 p_{1j} - \frac{\alpha^2}{(\alpha+1)^4} \eta_1^2 \sum_{j=1}^{N} x_{ij}^2 p_{1j}^2 \end{aligned} \tag{3.13}$$

Let $\zeta = \frac{\alpha}{(\alpha+1)^2}$ and $c_{i1} = -\frac{\alpha^2-\alpha}{(\alpha+1)^3} \eta_1 \sum_{j=1}^{N} x_{ij}^2 p_{1j} + \frac{\alpha^2}{(\alpha+1)^4} \eta_1^2 \sum_{j=1}^{N} x_{ij}^2 p_{1j}^2$, then (3.13) is $\zeta - c_{i1}$.

Next, we show r is approximately asymptotically same as $\frac{1}{\alpha+1}$. It is clear that r is the asymptotically same as $E(r)$. By the Taylor series approximation,

$$E(r) = E(\frac{1}{N} \sum_{j=1}^{N} y_j) = \frac{1}{N} \sum_{j=1}^{N} \frac{exp(\eta_0 + \eta_1 p_{j1})}{1 + exp(\eta_0 + \eta_1 p_{j1})} \approx \frac{1}{\alpha + 1}$$

.

By (i) and (ii), the approximated asymptotic distribution of S_i is

$$\frac{\zeta - c_{i1}}{\zeta}\chi_1^2(\frac{\zeta\eta_1 u_{i1}e_1}{2(\zeta - c_{i1})}),$$

and mean of S_i is

$$\frac{\zeta + \zeta^2\eta_1^2 u_{il}^2\lambda_1 - c_{i1}}{\zeta} \tag{3.14}$$

3.6 Tables and Figures

Table 3.1: Performance of the methods for 10,000 GWAS simulations. Values in the table represent type I error (for the null simulations) and power (for the alternative simulations) from LRT. The simulation setups are described in "Simulations and Real Data Analysis".

Trait	Simulation Type	Nominal Significance α	No Adjustment	Known Counfounding PCs	TW	EigenCorr1	EigenCorr2
Quantitative	NULL	0.05	0.1244	0.0494	0.0505	0.0502	0.0505
		10^{-2}	0.0507	0.0105	0.0103	0.0106	0.0111
	Alternative	10^{-2}	0.6294	0.7217	0.6972	0.7199	0.7196
		10^{-4}	0.2959	0.3717	0.3351	0.3694	0.3688
		10^{-6}	0.1161	0.1347	0.1115	0.1331	0.1324
Case Control	NULL	0.05	0.1642	0.0486	0.056	0.0488	0.0493
		10^{-2}	0.0765	0.0095	0.0120	0.0096	0.0098
	Alternative	10^{-2}	0.7291	0.8545	0.8483	0.8537	0.8526
		10^{-4}	0.4639	0.6358	0.6233	0.6334	0.6333
		10^{-6}	0.2663	0.4024	0.3911	0.3991	0.3983

Table 3.2: Performance of the methods for 10,000 GWAS simulations. Values in the table represent type I error (for the null simulations) and power (for the alternative simulations) from Score Test. The simulation setups are described in "Simulations and Real Data Analysis".

Trait	Simulation Type	Nominal Significance α	No Adjustment	Known Counfounding PCs	TW	EigenCorr1	EigenCorr2
Quantitative	NULL	0.05	0.1661	0.0501	0.0499	0.0508	0.0508
		10^{-2}	0.0782	0.0104	0.0106	0.0104	0.0105
	Alternative	10^{-2}	0.6871	0.7221	0.6983	0.7201	0.7191
		10^{-4}	0.3901	0.3743	0.3388	0.3708	0.3703
		10^{-6}	0.1877	0.1369	0.1155	0.1346	0.1334
Case Control	NULL	0.05	0.164	0.0485	0.0553	0.0488	0.0492
		10^{-2}	0.0763	0.0094	0.0120	0.0096	0.0098
	Alternative	10^{-2}	0.7276	0.8530	0.8464	0.8523	0.8515
		10^{-4}	0.4606	0.6327	0.6207	0.6308	0.6301
		10^{-6}	0.2626	0.3963	0.3850	0.3939	0.3935

Table 3.3: Genetic effect estimates for candidate SNPs from 10,000 simulations. Each entry shows the mean coefficient estimate, followed by the standard error in parentheses.

	true β	No Adjustment	Known Counfounding PCs	TW	EigenCorr1	EigenCorr2
Quantitative	0.15	0.148 (0.0799)	0.149 (0.0542)	0.149 (0.0565)	0.148 (0.0540)	0.148 (0.0543)
Case-Control	0.47	0.327 (0.1704)	0.472 (0.1367)	0.502 (0.1502)	0.471 (0.1366)	0.471 (0.1367)

Table 3.4: Performance of the methods for 10^9 GWAS simulations. Values in the table represent type I error from score test. The simulation setups are described in "Simulations and Real Data Analysis".

Trait	Nominal Significance α	No Adjustment	Known Counfounding PCs	TW	EigenCorr1	EigenCorr2
Quantitative	0.05	0.166	0.0500	0.0500	0.0499	0.0504
	10^{-4}	0.0162	9.95×10^{-5}	9.98×10^{-5}	9.85×10^{-5}	1.02×10^{-4}
	10^{-6}	0.00526	1.03×10^{-6}	1.03×10^{-6}	1.01×10^{-6}	1.09×10^{-6}
	10^{-7}	0.00327	9.76×10^{-8}	1.03×10^{-7}	9.25×10^{-8}	1.01×10^{-7}
Case Control	0.05	0.159	0.0483	0.0552	0.0483	0.0487
	10^{-4}	0.0144	9.42×10^{-5}	1.50×10^{-4}	9.40×10^{-5}	9.69×10^{-5}
	10^{-6}	0.00393	8.59×10^{-7}	1.81×10^{-6}	8.70×10^{-7}	9.05×10^{-7}
	10^{-7}	0.00218	7.70×10^{-8}	2.01×10^{-7}	7.78×10^{-8}	7.70×10^{-8}

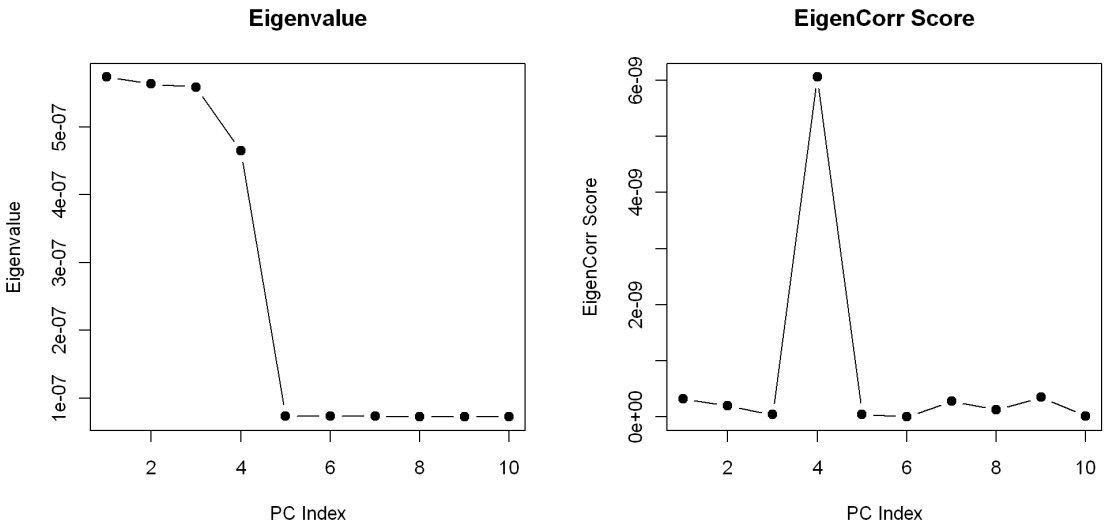

Figure 3.1: Illustration of the EigenCorr scores. The right panel presents the first 10 eigenvalues and the left panel presents the first 10 Eigencorr scores.

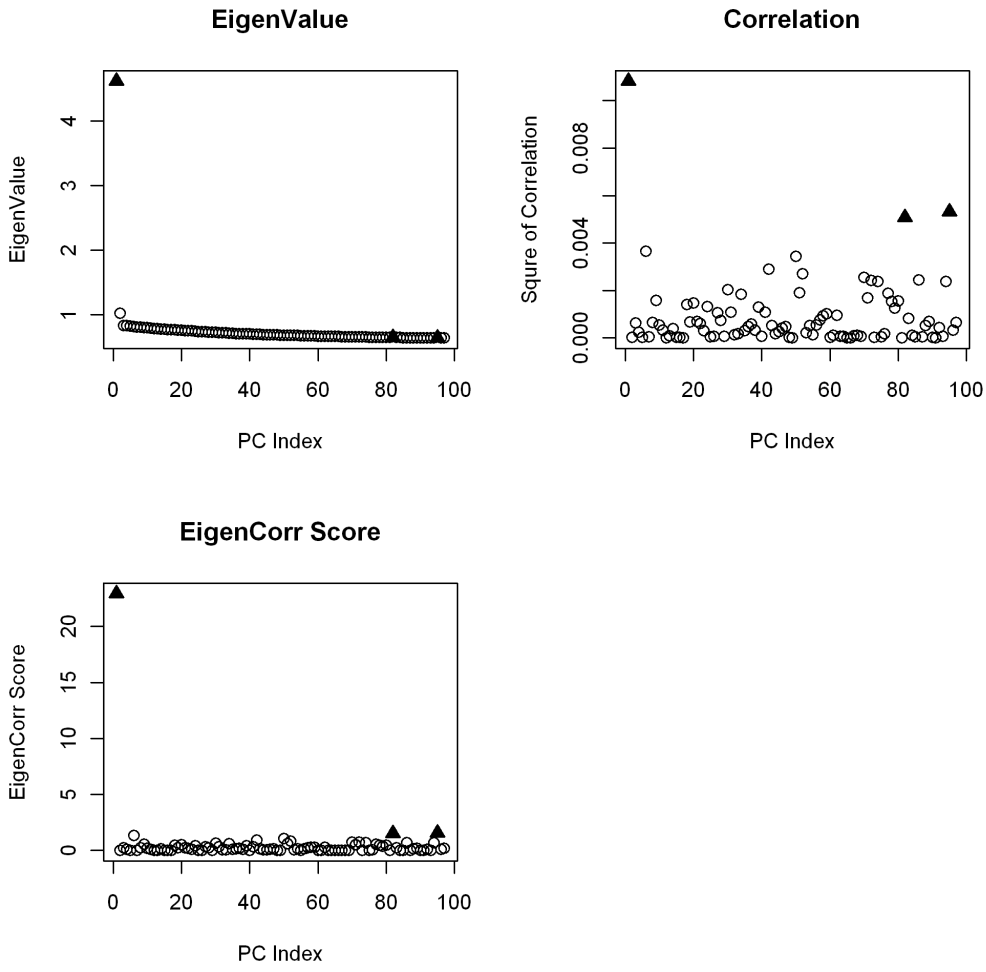

Figure 3.2: Eigenvalues, correlations and EigenCorr scores of PCs selected by the TW method, in schizophrenia dataset. Filled triangles represent PCs selected by either one of EigenCorr methods.

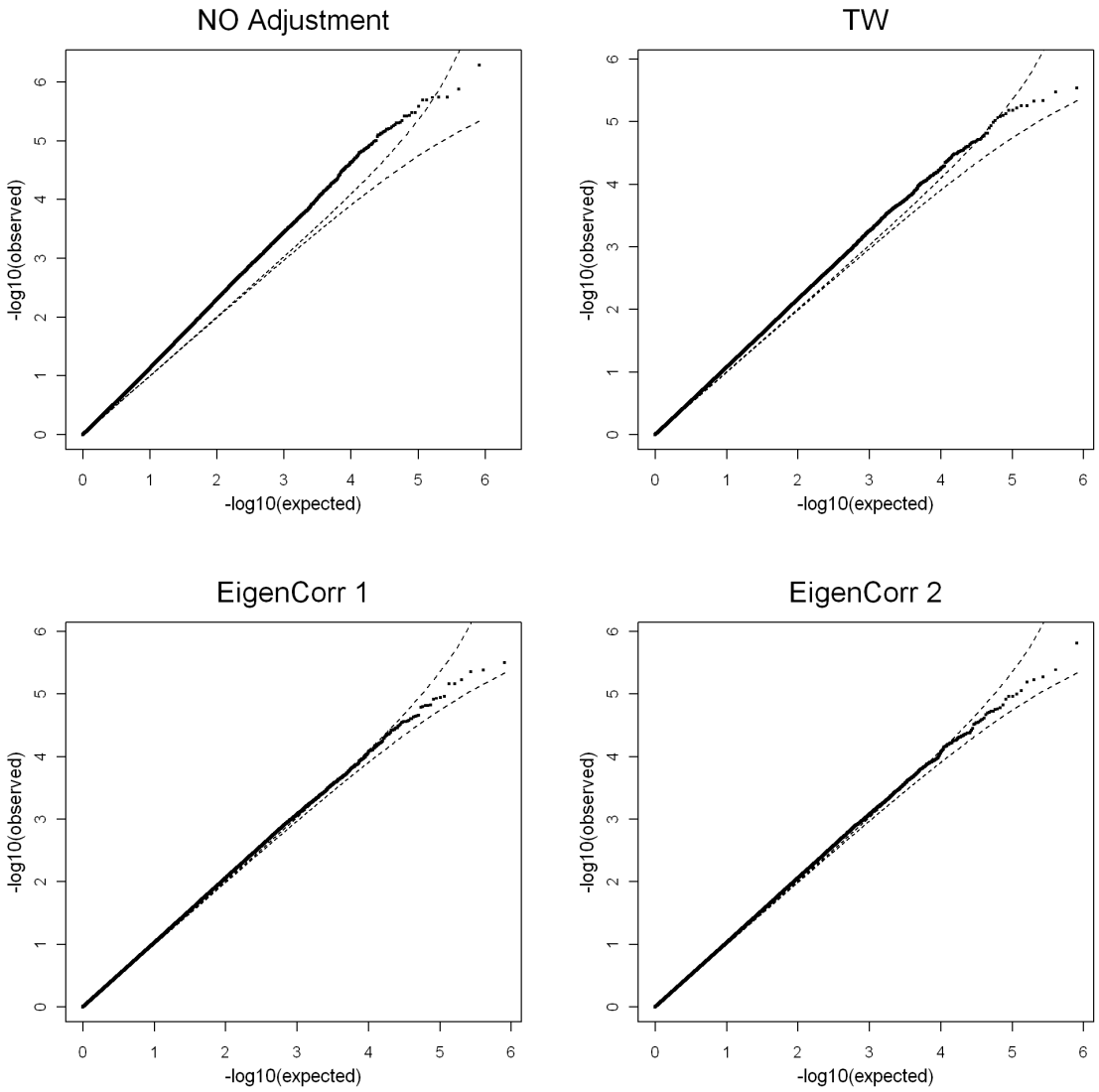

Figure 3.3: $-\log_{10}$ QQ plots of observed vs. expected p-values for the schizophrenia data. The dashed lines indicate 95% prediction bands.

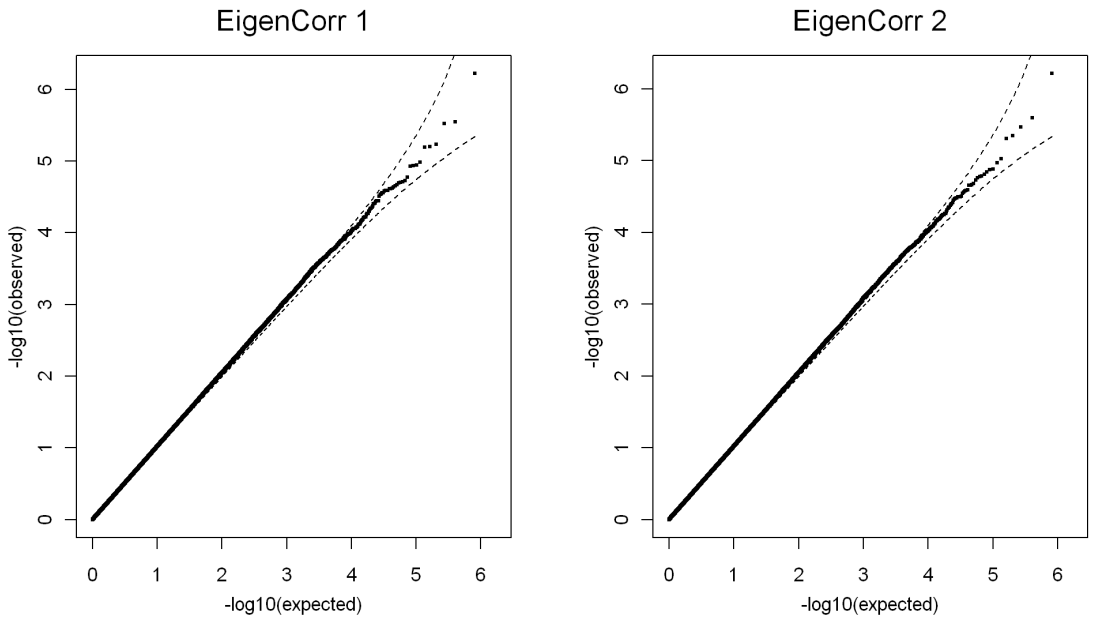

Figure 3.4: $-\log_{10}$ QQ plots of observed vs. expected p-values for the schizophrenia data. PCs were computed without SNP thinning and outlier exclusion. The dashed lines indicate 95% prediction bands.

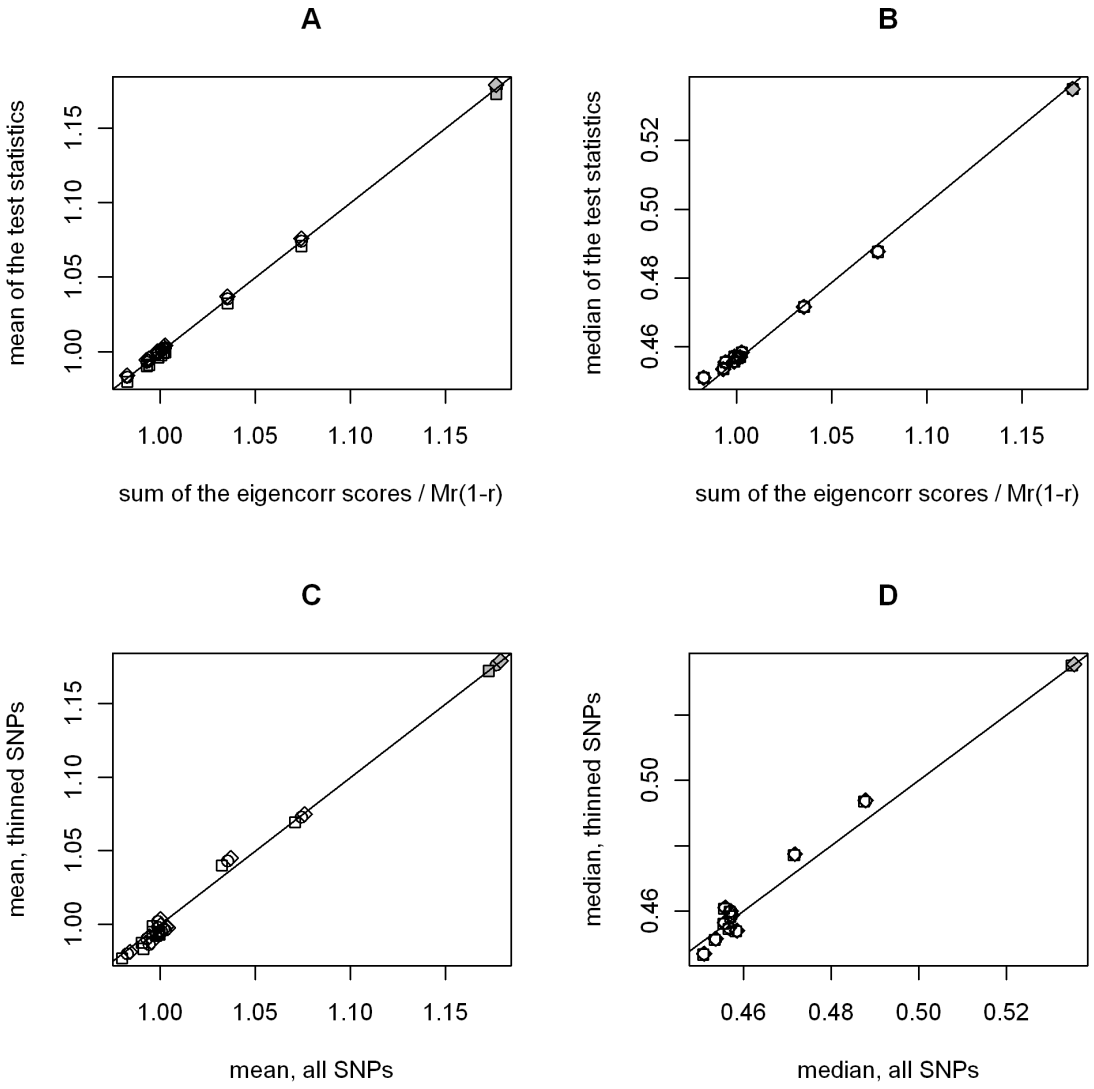

Figure 3.5: Comparison among score(circle), Wald(square), and likelihood ratio (diamond) test statistics vs. EigenCorr scores, based on 10 permuted outcomes (no color) and the real data (gray color). Panel A shows the mean of each test statistics vs. the appropriately scaled sum of EigenCorr scores. Note the good agreement for both permuted and real data. Panel B shows the median test statistics vs. the scaled EigenCorr scores, and again shows agreement with the theoretical line with slope 0.456. Panel C and D show mean and median test statistics of all SNPs vs. thinned SNPs.

Chapter 4

Convergence and Prediction of Principal Component Scores for High-Dimensional Matrices

A number of settings arise in which it is of interest to predict Principal Component (PC) scores for new observations using data from an initial sample. In this paper, we demonstrate that naive approaches to PC score prediction can be substantially biased towards 0 in the analysis of large matrices. This phenomenon is largely related to known inconsistency results for sample eigenvalues and eigenvectors as both dimensions of the matrix increase. For the spiked eigenvalue model for random matrices, we expand the generality of these results, and propose bias-adjusted PC score prediction. In addition, we compute the asymptotic correlation coefficient between PC scores from sample and population eigenvectors. Simulation and real data examples from the genetics literature show the improved bias and numerical properties of our estimators.

4.1 Introduction

Principal component analysis (PCA) (Jolliffe, 2002) is one of the leading statistical tools for analyzing multivariate data. It is especially popular in genetics/genomics, medical imaging, and chemometrics studies where high-dimensional data is common. PCA is typically used as a dimension reduction tool. A small number of top ranked principal component (PC) scores are computed by projecting data onto spaces spanned by the eigenvectors of sample covariance matrix, and are used to summarize data characteristics that contribute most to data variation. These PC scores can be subsequently used for data exploration and/or model predictions. For example, in genome-wide association studies (GWAS), PC scores are used to estimate ancestries of study subjects and as covariates to adjust for population stratification (Price et al., 2006; Patterson, Price and Reich, 2006). In gene expression microarray studies, PC scores are used as synthetic "eigen-genes" or "meta-genes" intended to represent and discover gene expression patterns that might not be discernible from single-gene analysis (Wall, Rechtsteiner and Rocha, 2003).

Although PCA is widely applied in a number of settings, much of our theoretical understanding rests on a relatively small body of literature. Girshick (1936) introduced the idea that the eigenvectors of sample covariance matrix are maximum likelihood estimators. Here a key concept in a population view of PCA is that the data arise as p-variate values from a distinct set of n independent samples. Later, the asymptotic distribution of eigenvalues and eigenvectors of the sample covariance matrix (i.e., the sample eigenvalues and eigenvectors) were derived for the situation where n goes to infinity and p is fixed (Girshick, 1939; Anderson, 1963). With the development of modern high-throughput technologies, it is not uncommon to have data where p is comparable in size to n, or substantially larger. Under the assumption that p and n grow at the same rate, that is $p/n \to \gamma > 0$, there has been considerable effort to

establish convergence results for sample eigenvalues and eigenvectors (see review (Bai, 1999)). The convergence of the sample eigenvalues and eigenvectors under the "spiked population" model proposed by Johnstone (2001) has also been established(Baik and Silverstein, 2006; Paul, 2007; Nadler, 2008). For this model it is well known that the sample eigenvectors are not consistent estimators of the eigenvectors of population covariance (i.e., the population eigenvectors) (Johnstone and Lu, 2007; Paul, 2007; Nadler, 2008). Furthermore, Paul (2007) has derived the degree of discrepancy in terms of the angle between the sample and population eigenvectors, under Gaussian assumptions for $0 < \gamma < 1$. More recently, Nadler (2008) has extended the same result to the more general $\gamma > 0$ using a matrix perturbation approach.

These results have considerable potential practical utility in understanding the behavior of PC analysis and prediction in modern datasets, for which p may be large. The practical goals of this paper focus primarily on the prediction of PC scores for samples which were not included in the original PC analysis. For example, gene expression data of new breast cancer patients may be collected, and we might want to estimate their PC scores in order to classify their cancer sub-type. The recalculation of PCs using both new and old data might not be practical, e.g. if the application of PCs from gene expression is used as a diagnostic tool in clinical applications. For GWAS analysis, it is known that PC analysis which includes related individuals tends to generate spurious PC scores which do not reflect the true underlying population substructures. To overcome this problem, it is common practice to include only one individual per family/sibship in the initial PC analysis. Another example arises in cross-validation for PC regression, in which PC scores for the test set might be derived using PCA performed on the training set (Jackson, 2005). For all of these applications, the predicted PC scores for a new sample are usually estimated in the "naive" fashion, in which the data vector of the new sample is multiplied by the sample eigenvectors from the original PC

analysis. Indeed, there appears to be relatively little recognition in the genetics or data mining literature that this approach may lead to misleading conclusions.

For low dimensional data, where p is fixed as n increases or otherwise much smaller than n, the predicted PC scores are nearly unbiased and well-behaved. However, for high-dimensional data, particularly with $p > n$, they tend to be biased and shrunken towards 0. The following simple example of a stratified population with three strata illustrates the shrinkage phenomenon for predicted PC scores. We generated a training data set with $n = 100$ and $p = 5000$. Among the 100 samples, 50 are from stratum 1, 30 are from stratum 2 and the rest from stratum 3. For each stratum, we first created a p-dimensional mean vector $\boldsymbol{\mu}_k$ $(k = 1, 2, 3)$. Each element of each mean vector was created by drawing randomly with replacement from $\{-0.3, 0, 0.3\}$, and thereafter considered a fixed property of the stratum. Then for each sample from the kth stratum, its p covariates were simulated from the multivariate normal distribution $MVN(\boldsymbol{\mu}_k, 4\mathbf{I})$, where $\mathbf{I}$ is the $p \times p$ identity matrix. A test dataset with the same sample size and $\boldsymbol{\mu}_k$ vectors was also simulated. Figure 4.1 shows that the predicted PC scores for the test data are much closer to 0 compared to the scores from the training data. This shrinkage phenomenon may create a serious problem if the predicted PC scores are used to classify new test samples, perhaps by similarity to previous apparent clusters in the original data. In addition, the predicted PC scores may produce incorrect results if used for downstream analyses (e.g., as covariates in association analyses).

In this paper, we investigate the degree of shrinkage bias associated with the predicted PC scores, and then propose new bias-adjusted PC score estimates. As the shrinkage phenomenon is largely related to the limiting behavior of the sample eigenvectors, our first step is to describe the discrepancy between the sample and population eigenvectors. To achieve this purpose, we follow the assumption that p and n both are large. By applying and extending results from random matrix theory, we establish the

convergence of the sample eigenvalues and eigenvectors under the spiked population model. We generalize Theorem 4 of Paul (2007), which describes the asymptotic angle between sample and population eigenvectors, to non-Gaussian random variables for any $\gamma > 0$. We further derive the asymptotic angle between PC scores from sample eigenvectors and population eigenvectors, and the asymptotic shrinkage factor of the PC score predictions. Finally we construct estimators of the angles and the shrinkage factor. The theoretical results are presented in Section 4.2.2. In Setion 4.2.3, we extended our theoretical results to the case that p is substantially larger than n, and thus $p/n \to \infty$.

In section 4.3, we report simulations to assess the finite sample accuracy of the proposed asymptotic angle and shrinkage factor estimators. We also show the potential improvements in prediction accuracy for PC regression by using the bias adjusted PC scores. In Section 4.4, we apply our PC analysis to a real genome-wide association study, which demonstrates that the shrinkage phenomenon occurs in real studies and that adjustment is needed.

4.2 Materials and Methods

4.2.1 General Setting

Throughout this paper, we use ^{T} to denote matrix transpose, $\xrightarrow{p}$ to denote convergence in probability, and $\xrightarrow{a.s}$ to denote almost sure convergence. Let $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$, a $p \times p$ matrix with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, and $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_p]$, a $p \times p$ orthogonal matrix.

Define the $p \times n$ data matrix, $\mathbf{X}$ as $[\mathbf{x}_1, \ldots, \mathbf{x}_n]$, where $\mathbf{x}_j$ is the p-dimensional vector corresponding to the j^{th} sample. For the remainder of the paper, we assume the following:

Assumption 1. $\mathbf{X} = \mathbf{E}\mathbf{\Lambda}^{1/2}\mathbf{Z}$, where $\mathbf{Z} = \{z_{ij}\}$ is a $p \times n$ matrix whose elements z_{ij}s are i.i.d random variables with $E(z_{ij}) = 0, E(z_{ij}^2) = 1$ and $E(z_{ij}^4) < \infty$.

Although the z_{ij}s are i.i.d, Assumption 1 allows for very flexible covariance structures for $\mathbf{X}$, and thus the results of this paper are quite general. The population covariance matrix of $\mathbf{X}$ is $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$. The sample covariance matrix $\mathbf{S}$ equals

$$\mathbf{S} = \mathbf{X}\mathbf{X}^T/n = \mathbf{E}\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Lambda}^{1/2}\mathbf{E}^T/n.$$

The λ_ks are the underlying population eigenvalues. The spiked population model defined in (Johnstone, 2001) assumes that all the population eigenvalues are 1, except the first m eigenvalues. That is, $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m > \lambda_{m+1} = \cdots = \lambda_p = 1$. The spectral decomposition of the sample covariance matrix is

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{U}^T,$$

where $\mathbf{D} = diag(d_1, d_2, \ldots, d_p)$ is a diagonal matrix of the ordered sample eigenvalues and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ is the corresponding $p \times p$ sample eigenvector matrix. Then the PC score matrix is $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n]$, where $\mathbf{p}_v^T = \mathbf{u}_v^T\mathbf{X}$ is the vth sample PC score. For a new observation $\mathbf{x}_{new}$, its predicted PC score is similarly defined as $\mathbf{U}^T\mathbf{x}_{new}$ with the vth (PC) score equal to $q_v = \mathbf{u}_v^T\mathbf{x}_{new}$.

4.2.2 When $p/n \to \gamma < \infty$

Sample Eigenvalues and Eigenvectors

Under the classical setting of fixed p, it is well known that the sample eigenvalues and eigenvectors are consistent estimators of the corresponding population eigenvalues and

eigenvectors (Anderson, 2003). Under the "large p, large n" framework, however, the consistency is not guaranteed. The following two lemmas summarize and extend some known convergence results.

Lemma 1 *Let $p/n \to \gamma \geq 0$ as $n \to \infty$.*

i) When $\gamma = 0$,

$$d_v \overset{a.s}{\to} \begin{cases} \lambda_v, & \text{for } v \leq m \\ 1, & \text{for } v > m; \end{cases} \tag{4.1}$$

ii) When $\gamma > 0$,

$$d_v \overset{a.s}{\to} \begin{cases} \rho(\lambda_v), & \text{for } v \leq k \\ (1+\sqrt{\gamma})^2, & \text{for } v = k+1, \end{cases} \tag{4.2}$$

where k is the number of λ_v greater than $1+\sqrt{\gamma}$, and $\rho(x) = x(1+\gamma/(x-1))$.

The result in ii) is due to Baik and Silverstein (2006), while the proof of i) can be found in section (4.6.2). The result in i) shows that when $\gamma = 0$, the sample eigenvalues converge to the corresponding population eigenvalues, which is consistent with the classical PC result where p is fixed. The result in ii) shows that for any non-zero γ, d_v is no longer a consistent estimator of λ_v. However, a consistent estimator of λ_v can be constructed from Equation (4.2). Define

$$\rho^{-1}(d) = \frac{d+1-\gamma+\sqrt{(d+1-\gamma)^2-4d}}{2}.$$

Then $\rho^{-1}(d_v)$ is a consistent estimator of λ_v when $\lambda_v > 1+\sqrt{\gamma}$. Furthermore, Baik, Ben Arous and Peche (2005) have shown the $\sqrt{n}$-consistency of d_v to $\rho(\lambda_v)$, and Bai and Yao (2008) have shown that d_v is asymptotically normal.

Lemma 2 *Suppose $p/n \to \gamma \geq 0$ as $n \to \infty$. Let $< ., . >$ be an inner product between two vectors. Under the assumption of multiplicity one,*

i) if $0 < \gamma < 1$*, and the* $z_{ij}s$ *follow the standard normal distribution, then*

$$| < \mathbf{e}_v, \mathbf{u}_v > | \xrightarrow{a.s} \begin{cases} \phi(\lambda_v), & \text{if } \lambda_v > 1 + \sqrt{\gamma} \\ 0, & \text{if } 1 < \lambda_v \leq 1 + \sqrt{\gamma} \end{cases} \tag{4.3}$$

ii) removing the normal assumption on the $z_{ij}s$*, the following weaker convergence result holds for all* $\gamma \geq 0$

$$| < \mathbf{e}_v, \mathbf{u}_v > | \xrightarrow{p} \begin{cases} \phi(\lambda_v), & \text{if } \lambda_v > 1 + \sqrt{\gamma} \\ 0, & \text{if } 1 < \lambda_v \leq 1 + \sqrt{\gamma}. \end{cases} \tag{4.4}$$

Here $\phi(x) = \sqrt{(1 - \frac{\gamma}{(x-1)^2})/(1 + \frac{\gamma}{x-1})}$.

The inner product between unit vectors is the cosine angle between these two. Thus, Lemma 2 shows the convergence of the angle between population and sample eigenvectors. For i), Paul (2007) proved it for $\gamma < 1$; while Nadler (2008) obtained the same conclusion for $\gamma > 0$ using the matrix perturbation approach under the Gaussian random noise model. We relax the Gaussian assumption on z and prove ii) for $\gamma \geq 0$ in section 4.6.3. The result of ii) is general enough for the application of PCA to, for example, genome-wide association mapping, where each entry of $\mathbf{X}$ is a standardized variable of SNP genotypes, which are typically coded as $\{0, 1, 2\}$, corresponding to discrete genotypes.

Sample and Predicted PC Scores

In this section, we first discuss convergence of the sample PC scores, which forms the basis for the investigation of the shrinkage phenomenon of the predicted PC scores. For the sample PC scores, we have

Theorem 2 *Let* $\mathbf{g}_v^T = \mathbf{e}_v^T\mathbf{X}/\sqrt{\mathbf{n}\lambda_v}$*, the normalized* v^{th} *PC score derived from a corresponding population eigenvector,* $\mathbf{e}_v$*, and* $\tilde{\mathbf{p}}_v = \mathbf{p}_v/\sqrt{d_v}$*, the normalized* v^{th} *sample PC score. Suppose* $p/n \to \gamma \geq 0$ *as* $n \to \infty$*. Under the multiplicity one assumption,*

$$| < \mathbf{g}_v, \tilde{\mathbf{p}}_v > | \xrightarrow{p} \begin{cases} \sqrt{1 - \frac{\gamma}{(\lambda_v - 1)^2}}, & \text{if } \lambda_v > 1 + \sqrt{\gamma} \\ 0, & \text{if } 1 < \lambda_v \leq 1 + \sqrt{\gamma}. \end{cases} \tag{4.5}$$

The proof can be found in section 4.6.4. In PC analysis, the sample PC scores are typically used to estimate certain latent variables (largely the PC scores from population eigenvectors) that represent the underlying data structures. The above result allows us to quantify the accuracy of the sample PC scores. Note that here $< \mathbf{g}_v, \tilde{\mathbf{p}}_v >$ is the correlation coefficient between $\mathbf{g}_v$ and $\tilde{\mathbf{p}}_v$. Compared to Equation (4.3) in Lemma 2, the angle between the PC scores is smaller than the angle between their corresponding eigenvectors.

Before we formally derive the asymptotic shrinkage factor for the predicted PC scores, we first describe in mathematical terms the shrinkage phenomenon that was demonstrated in the Introduction. Note that the first population eigenvector $\mathbf{e}_1$ satisfies

$$\mathbf{e}_1 = \underset{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1}{\operatorname{argmax}} E((\mathbf{a}^T\mathbf{x})^2)$$

for a random vector $\mathbf{x}$ that follows the same distribution of the $\mathbf{x}_j$s. For the data matrix $\mathbf{X}$, its first sample eigenvector $\mathbf{u}_1$ satisfies

$$\mathbf{u}_1 = \underset{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1}{\operatorname{argmax}} \sum_{j=1}^{n} (\mathbf{a}^T\mathbf{x}_j)^2.$$

Assuming that $\mathbf{u}_1$ and the new sample $\mathbf{x}_{new}$ are independent of each other, we have

$$\begin{aligned} E((\mathbf{u}_1^T\mathbf{x}_{new})^2)) &= E(E(\mathbf{u}_1^T\mathbf{x}_{new}\mathbf{x}_{new}^T\mathbf{u}_1^T|\mathbf{u}_1)) = E(\mathbf{u}_1^T E(\mathbf{x}_{new}\mathbf{x}_{new}^T)\mathbf{u}_1^T) \\ &= E(\mathbf{u}_1^T\boldsymbol{\Sigma}\mathbf{u}_1^T) \leq \mathbf{e}_1^T\boldsymbol{\Sigma}\mathbf{e}_1 = E((\mathbf{e}_1^T\mathbf{x}_{new})^2). \end{aligned} \tag{4.6}$$

Since the $\mathbf{u}_1^T\mathbf{x}_j$s $(j = 1, \ldots, n)$ follow the same distribution,

$$nE((\mathbf{e}_1^T\mathbf{x}_j)^2) = E(\sum_{j=1}^{n}(\mathbf{e}_1^T\mathbf{x}_j)^2) \leq E(\sum_{j=1}^{n}(\mathbf{u}_1^T\mathbf{x}_j)^2) = nE((\mathbf{u}_1^T\mathbf{x}_j)^2). \tag{4.7}$$

By (4.6) and (4.7), we can show that

$$E((\mathbf{u}_1^T\mathbf{x}_{new})^2) \leq E((\mathbf{e}_1^T\mathbf{x}_{new})^2) = E((\mathbf{e}_1^T\mathbf{x}_j)^2) \leq E((\mathbf{u}_1^T\mathbf{x}_j)^2),$$

which demonstrates the shrinkage feature of the predicted PC scores. The amount of the shrinkage, or the asymptotic shrinkage factor, is given by the following theorem:

Theorem 3 *Suppose* $p/n \to \gamma \geq 0$ *as* $n \to \infty$, $\lambda_v > 1 + \sqrt{\gamma}$. *Under the multiplicity one assumption,*

$$\sqrt{\frac{E(q_v^2)}{E(p_{vj}^2)}} \overset{n\to\infty}{\to} \frac{\lambda_v - 1}{\lambda_v + \gamma - 1} \tag{4.8}$$

where p_{vj} *is the* j^{th} *element of* $\mathbf{p}_v$.

The proof is given in section 4.6.5. We call $(\lambda_v - 1)/(\lambda_v + \gamma - 1)$, the (asymptotic) shrinkage factor for a new subject. As shown, the shrinkage factor is smaller than 1 if $\gamma > 0$. Quite sensibly, it is a decreasing function of γ and an increasing function of λ_v. The bias of the predicted PC score can be potentially large for those high dimensional data where p is substantially greater than n, and/or for the data with relatively minor underlying structures where λ_v is small.

Rescaling of sample eigenvalues

The previous theorems are based on the assumption that all except the top m eigenvalues are equal to 1. Even under the spiked eigenvalue model, some rescaling of the sample eigenvalues may be necessary with real data.

For a given data, let its ordered population eigenvalues $\mathbf{\Lambda}^* = \{\zeta\lambda_1, \ldots, \zeta\lambda_m, \zeta, \ldots, \zeta\}$, where $\zeta \neq 1$, and its corresponding sample eigenvalues $\mathbf{D}^* = \{d_1^*, \ldots, d_n^*\}$. We can show that Equations (4.4), (4.8), and (4.5) still hold under such circumstances. However, $\rho^{-1}(d_v^*)$ is no longer a consistent estimator of λ_v, because

$$d_v^* \overset{a.s}{\to} \zeta\lambda_v(1 + \frac{\gamma}{\lambda_v - 1}) = \zeta\rho(\lambda_v).$$

To address this issue, Baik and Silverstein (2006) have proposed a simple approach to estimate ζ. In their method, the top significant large sample eigenvalues are first separated from the other grouped sample eigenvalues. Then ζ is estimated as the ratio between the average of the grouped sample eigenvalues and the mean determined by the Marchenko-Pastur law (Marčenko and Pastur, 1967). To separate the eigenvalues, they have suggested to use a screeplot of the percent variance versus component number. However, for real data, we may not be able to clearly separate the sample eigenvalues in such a manner and readily apply the approach. Thus we need an automated method which does not require a clear separation of the sample eigenvalues.

The expectation of the sum of the sample eigenvalues when $\zeta = 1$ is

$$E(\sum_{v=1}^{p} d_v) = E(trace(\mathbf{S})) = trace(E(\mathbf{S})) = trace(\mathbf{\Sigma}) = \sum_{v=1}^{p} \lambda_v,$$

Thus, the sum of the rescaled eigenvalues is expected to be close to $(\sum_{v=1}^{m} \lambda_v + p - m)$. Let $r_v = d_v^* / \left(\sum_{v=1}^{p} d_v^*\right)$ and $\hat{d}_v$ be a properly rescaled eigenvalue, then $\hat{d}_v$ should be very

close to $r_v(\sum_{v=1}^{m}\lambda_v + p - m)$. Note that $p/(\sum_{v=1}^{m}\lambda_v + p - m) \to 1$ for fixed m and λ_v. Thus pr_v is a properly adjusted eigenvalue. However, for finite n and p, the difference between p and $(\sum_{v=1}^{m}\lambda_v + p - m)$ can be substantial, especially when the first several λ_vs are considerably larger than 1. To reduce this difference, we propose a novel method which iteratively estimates the $(\sum_{v=1}^{m}\lambda_v + p - m)$ and $\hat{d}_v$.

1. Initially set $\hat{d}_{v,0} = pr_v$
2. For the l^{th} iteration, set $\hat{\lambda}_{v,l} = \rho^{-1}(\hat{d}_{v,l-1})$ for $\hat{d}_{v,l-1} > (1+\sqrt{\gamma})^2$, and $\hat{\lambda}_{v,l} = 1$ for $\hat{d}_{v,l-1} \leq (1+\sqrt{\gamma})^2$. Define k_l as the number of $\hat{\lambda}_{v,l}$s that are greater than 1, and let

$$\hat{d}_{v,l} = \left(\sum_{v=1}^{k_l}\hat{\lambda}_{v,l} + p - k_l\right) r_v.$$

3. If $\sum_{v=1}^{k_l}\hat{\lambda}_{v,l} + p - k_l$ converges, let

$$\hat{d}_v = \hat{d}_{v,l}$$

and stop. Otherwise, go to step 2.

The consistency of $\hat{d}_v$ to $\rho(\lambda_v)$ is shown in the following theorem.

Theorem 4 *Let $\hat{d}_v$ be the rescaled sample eigenvalue from the proposed algorithm. Then, for $\lambda_v > 1+\sqrt{\gamma}$ with multiplicity one,*

$$\hat{d}_v \xrightarrow{p} \rho(\lambda_v)$$

Since $\rho^{-1}(\hat{d}_v) \xrightarrow{p} \lambda_v$, $\phi(\rho^{-1}(\hat{d}_v))^2$ is a consistent estimator of $\phi(\lambda_v)^2$. Combining this

fact with Theorems 1 and 2, we can obtain the bias adjusted PC score q_v^*

$$q_v^* = q_v \frac{\rho^{-1}(\hat{d}_v) + \gamma - 1}{\rho^{-1}(\hat{d}_v) - 1}$$

and the asymptotic correlation coefficient between $\mathbf{g}_v$ and $\tilde{\mathbf{p}}_v$

$$\sqrt{(1 - \frac{\gamma}{(\rho^{-1}(\hat{d}_v) - 1)^2})}.$$

4.2.3 When $p/n \to \infty$

Convergence of sample eigenvalues, eigenvectors and PC scores

One of the main assumptions of previous theoretical results is the same increment rate of p and n, and thus $p/n \to \gamma < \infty$. For many modern data, however, p is substantially larger than n. As a result, this same increment rate assumption may not be satisfied for those ultra high dimensional data. In this section, we investigate asymptotic behaviors of sample eigenvalues, eigenvectors and PC scores with $\gamma \to \infty$.

Define $\hat{\gamma}_{p,n} = p/n$ and assume $\hat{\gamma}_{p,n} \to \infty$ as both p and n grow to ∞. In previous sections, λ_v is assumed to be fixed. However, it is obvious that if λ_v is fixed under new asymptotic setting, true signal would be overwhelmed by noise. Thus, we allow λ_v increases as p increases. In particular, we set

$$\lambda_{v,p} = c_{v,p,n}\hat{\gamma}_{p,n},$$

for $v \leq m$, where $c_{v,p,n}$ can go to ∞ and to 0. For notational simplicity, we suppress subscripts p and n.

Before introducing main results, we define symbols for further use. Suppose a_p and b_p are two sequences. We denote $a_p \asymp b_p$ as $a_p = O(b_p)$ and $b_p = O(a_p)$, $a_p \gg b_p$ as

$b_p/a_p = o(1)$, and $a_p \ll b_p$ as $a_p/b_p = o(1)$. Below theorem shows the convergence of sample eigenvalues under the new setting.

Theorem 5 *Suppose $\hat{\gamma} \to \infty$ as both p and $n \to \infty$. Let $\lambda_v = c_v\hat{\gamma}$ for $v \leq m$ with $c_1 \asymp \ldots \asymp c_m$, and $\lambda_v = 1$ for $v > m$. z_{ij}s satisfy Assumption 1, then with multiplicity 1,*

i) When c_v is bounded away from zero,

$$\frac{d_v}{\lambda_v} - \frac{c_v + 1}{c_v} \xrightarrow{a.s} 0, \quad \text{for } v \leq m$$

$$\frac{d_v}{\hat{\gamma}} \xrightarrow{a.s} 1, \quad \text{for } v > m$$

ii) When $c_v = o(1)$,

$$\frac{d_v}{\hat{\gamma}} \xrightarrow{a.s} 1, \quad \text{for all } v$$

A proof can be found in Section 4.6.7. Theorem 5 shows that spiked eigenvalues are separated from the bulk when c_v is bigger than zero. If $c_v = o(1)$, we cannot recover the signal of spiked population eigenvalues from sample eigenvalues. Theorem 5 also presents that sample eigenvalues are not consistent to the population eigenvalues for the finite c_v. This conclusion coincides with the previous results based on the finite γ. When $c_v \to \infty$, and then $(c_v + 1)/(c_v) \to 1$, which indicates the consistency of sample eigenvalues to population eigenvalues.

Theorem 6 *Suppose $\hat{\gamma} \to \infty$ as both p and $n \to \infty$. Let $\lambda_v = c_v\hat{\gamma}$ for $v \leq m$ with $c_1 \asymp \ldots \asymp c_m$, and $\lambda_v = 1$ for $v > m$. z_{ij}s satisfy Assumption 1, then with multiplicity 1,*

i) When c_v is bounded away from zero,

$$| < \mathbf{e}_v, \mathbf{u}_v > | - \sqrt{\frac{c_v}{c_v + 1}} \xrightarrow{p} 0, \quad \text{for } v \leq m$$

ii) When $c_v = o(1)$,

$$| < \mathbf{e}_1, \mathbf{u}_1 > | \xrightarrow{p} 0$$

A proof is given in Section 4.6.8. This theorem shows that the consistency or inconsistency of sample eigenvectors are determined by the increment rate of c_v. If $c_v \to \infty$, and then $c_v/(c_v+1) \to 1$, which means the v^{th} sample eigenvector is consistent to the corresponding population eigenvector. When c_v is finite but bigger than zero, the v^{th} sample eigenvector is not consistent and is not perpendicular to the corresponding population eigenvector. In the case of $c_v = o(1)$, v^{th} sample eigenvector is perpendicular to the corresponding population eigenvector. It can be notified that the case of the finite c_v corresponds to the previous results.

Theorem 7 *Suppose* $\hat{\gamma} \to \infty$ *as both* p *and* $n \to \infty$. *Let* $\lambda_v = c_v\hat{\gamma}$ *for* $v \leq m$ *with* $c_1 \asymp \ldots \asymp c_m$, *and* $\lambda_v = 1$ *for* $v > m$. z_{ij}*s satisfy Assumption 1, and* λ_v *has multiplicity 1. When* c_v *is bounded away from zero,*

$$| < \mathbf{g}_v, \tilde{\mathbf{p}}_v > | \xrightarrow{p} 1,$$

where $v \leq m$.

See a proof of Theorem 7 in Section 4.6.9. One striking feature is that the inner product between $\mathbf{g}_v$ and $\tilde{\mathbf{p}}_v$ converges to 1 when c_v is larger than zero, although the corresponding sample eigenvector is not consistent. As shown in Theorem 5, spiked eigenvalues can be identified whenever c_v is bigger than zero. Combining Theorem 5 and Theorem 7, we can conclude that $\tilde{\mathbf{p}}_v$ can accurately estimate $\mathbf{g}_v$ whenever its corresponding sample eigenvalue is separated from the bulk. This very interesting result presents why PCA is very successful in many high dimensional data.

Theorem 8 *Suppose $\hat{\gamma} \to \infty$ as both p and $n \to \infty$. Let $\lambda_v = c_v\hat{\gamma}$ for $v \leq m$ with $c_1 \asymp \ldots \asymp c_m$, and $\lambda_v = 1$ for $v > m$. z_{ij}s satisfy Assumption 1, and λ_v has multiplicity 1. When c_v is bounded away from zero,*

$$\sqrt{\frac{E(q_v^2)}{E(p_{vi}^2)}} - \frac{c_v}{c_v + 1} \xrightarrow{p} 0,$$

where $v \leq m$.

A proof is given in Section 4.6.10. This theorem shows that the shrinkage bias occurs in the ultra-high dimensional scenario.

Relation to the finite γ asymptotics

Under finite γ asymptotics, spiked eigenvalues are separated from the bulk, and an angle of sample and population eigenvectors is bigger than zero but smaller than $\pi/2$ when the corresponding population eigenvalue is bigger than $1+\sqrt{\gamma}$. If λ_v is increasing as $p \to \infty$, it can be easily shown that the $d_v/\lambda_v \to 1$ and $< e_v, d_v > \to 1$. These observations suggest that the fixed λ_v is equivalent to $c_v = O(1)$ and the increasing λ_v is equivalent to $c_v \gg O(1)$.

The asymptotic results in Section 4.2.3 can be derived from the finite γ asymptotics, by substituting λ_v to $c_v\hat{\gamma}$ and increasing $\hat{\gamma}$ to ∞. For example,

$$\frac{d_v}{\hat{\gamma}} \to \frac{\lambda_v}{\hat{\gamma}}(1 + \frac{\hat{\gamma}}{\lambda_v - 1}),$$

for fixed $\hat{\gamma}$. Substitute λ_v to $c_v\hat{\gamma}$, then

$$\frac{d_v}{\hat{\gamma}} \to c_v(1 + \frac{\hat{\gamma}}{c_v\hat{\gamma} - 1}) \approx \frac{c_v}{1 + 1/c_v},$$

for large $\hat{\gamma}$.

The same conclusion can be made for an angle between sample and population eigenvectors, an angle between PC scores from sample and population eigenvectors, and a shrinkage factor of predicted PC score. It shows that the asymptotic results of sample eigenvalues, eigenvectors, and PC scores in both finite and infinite γ cases are nearly identical for large γ, which justifies using the results from Section 4.2.2 to ultra-high dimensional data.

4.3 Simulation

First, we applied our bias adjustment process to the simulated data described in the Introduction. Our estimated asymptotic shrinkage factors are 0.465 and 0.329 for the first and second PC scores, respectively. The scatter plot of the top two bias adjusted PC scores is given in Figure 4.2. After the bias adjustment, the predicted PC scores of the test data are comparable to those of the training data. This indicates that our method is effective in correcting for the shrinkage bias.

Next, we conducted a new simulation to check the accuracy of our estimators. For the jth sample ($j = 1, \ldots, n$), its i^{th} variable was generated as

$$x_{ij} = \begin{cases} \lambda_1 z_{ij} & i = 1 \\ \lambda_2 z_{ij} & i = 2 \\ z_{ij} & i > 2 \end{cases}$$

where $\lambda_1 > \lambda_2 > 1$ and $z_{ij} \sim N(0, 2^2)$. Under this setting, λ_1 and λ_2 are the first and the second population eigenvalues. The first and second population eigenvectors are $e_1 = \{1, 0, \ldots, 0\}$ and $e_2 = \{0, 1, 0, \ldots, 0\}$ respectively. We set the standard deviation of z_{ij} to 2 instead of 1, which allows us to test whether the rescaling procedure works properly. We tried different values of γ and n, but set λ_1 and λ_2 to $4(1 + \sqrt{\gamma})$ and

$2(1+\sqrt{\gamma})$, respectively.

We split the simulated samples into test and training sets, each with n samples. We first estimated the asymptotic shrinkage factor based on the training samples. We then calculated the predicted PC scores on the test samples. To assess the accuracy of shrinkage factor estimator for each PC, we empirically estimated the shrinkage factor by the ratio of the mean predicted PC scores of the test samples to the mean PC scores of the training samples. That is, for the vth PC, the empirical shrinkage factor is estimated by $\sqrt{\sum_{i=1}^{n} q_{vi}^2 / \sum_{k=1}^{n} p_{vk}^2}$. On the training samples, we also estimated the empirical angle between the sample and (known) population eigenvectors, as well as the empirical angle between PC scores from sample and population eigenvectors. The asymptotic theoretical estimates were also calculated. Tables 4.1 and 4.2 summarize the simulation results. Our asymptotic estimators provide accurate estimates for the angles and the shrinkage factor.

Finally, we conducted simulation to demonstrate an application of the bias adjusted PC scores in PC regression. PC regression has been widely used in microarray gene-expression studies (Bovelstad et al., 2007). In this simulation, we let $p = 5,000$, and our set up is very similar to the first simulation of Bair et al. (2006). Let x_{ij} denote the gene expression level of the ith gene for the jth subject. We generated each x_{ij} according to

$$x_{ij} = \begin{cases} 3+\epsilon & i \le g, j \le n/2 \\ 4+\epsilon & i \le g, j > n/2 \\ 3.5+\epsilon & i > g \end{cases}$$

and the outcome variable y_j as

$$y_j = \frac{2}{g}\sum_{i=1}^{g} x_{ij} + \epsilon_y,$$

where n is the number of samples, g is the number of genes that are differentially

expressed and associated with the phenotype, $\epsilon \sim N(0, 2^2)$ and $\epsilon_y \sim N(0, 1)$. A total of eight different combinations of n and g were simulated. For the training data, we fit the PC regression with the first PC as the covariate and computed the mean square error (MSE). For the test samples with the same configuration of the training samples, we applied the PC model built on the training data to predict the phenotypes using the un-adjusted and adjusted PC scores. The results are presented in Table 4.3. We see that the MSE of the test set without bias adjustment is appreciably higher than that of the test set with bias adjustment, and the MSE of the test set with bias adjustment is comparable with the MSE of the training set.

4.4 Real data example

Here we demonstrate that the shrinkage phenomenon appears in real data, and can be adjusted by our method. For this purpose, genetic data on samples from unrelated individuals in the Phase 3 HapMap study [http://hapmap.ncbi.nlm.nih.gov/] were used. HapMap is a dense genotyping study designed to elucidate population genetic differences. The genetic data are discrete, assuming the values 0, 1, or 2 at each genomic marker (also known as SNPs) for each individual. Data from CEU individuals (of northern and western European ancestry) were compared with data from TSI individuals (Toscani individuals from Italy, representing southern European ancestry).

Some initial data trimming steps are standard in genetic analysis. We first removed apparently related samples, and removed genomic markers with more than a 10% missing rate, and those with frequency less than 0.01 for the minor genetic allele. To avoid spurious PC results, we further pruned out SNPs that are in high linkage disequlibrium (LD) (Fellay et al., 2007). Lastly, we excluded 7 samples with PC scores greater than 6 standard deviations away from the mean of at least one of the top significant PCs (i.e., with Tracy-Widom (TW) Test p-value < 0.01) (Price et al., 2006; Patterson, Price and

Reich, 2006). The final dataset contained 178 samples (101 CEU, 77 TSI) and 100,183 markers. We mean-centered and variance-standardized the genotypes for each marker (Price et al., 2006). The screeplot of the sample eigenvalues is presented in Figure 4.3. The first eigenvalue is substantially larger than the rest of the eigenvalues, although the TW test actually identifies two significant PCs. Figure 4.3 suggests that our data approximately satisfies the spiked eigenvalue assumption.

We estimated the asymptotic shrinkage factor and compared it with the following jackknife-based shrinkage factor estimate. For the first PC, we first computed the scores of all samples. Next, we removed one sample at a time and computed the (unadjusted) predicted PC score. We then calculated the jackknife estimate as the square root of the ratio of the means of the sample PC score and the predicted PC score. The jackknife shrinkage factor estimate is 0.319, which is close to our asymptotic estimate 0.325. Figure 4.4 shows the PC scores from the whole sample, the predicted PC score of an illustrative excluded sample, and its bias-adjusted predicted score. Clearly, the predicted PC score without adjustment is very biased towards zero, while the bias adjusted PC score is not.

4.5 Discussion and conclusions

In this paper we have identified and explored the shrinkage phenomenon of the predicted PC scores, and have developed a novel method to adjust these quantities. We also have constructed the asymptotic estimator of correlation coefficient between PC scores from population eigenvectors and sample eigenvectors. In simulation experiments and real data analysis, we have demonstrated the accuracy of our estimates, and the capability to increase prediction accuracy in PC regression by adopting shrinkage bias adjustment. For achieving these, we consider asymptotics in the large p, large n framework, under the spiked population model.

Although the results from the spiked model are useful, it is likely that observed data has more structure than allowed by the model. Recently, several methods have been suggested to estimate population eigenvalues under more general scenarios (El Karoui, 2008; Rao et al., 2008). However, no analogous results are available for the eigenvectors. In data analysis, jackknife estimators, as demonstrated in the real data analysis section, can be used. However, resampling approaches are very computationally intensive, and it remains of interest to establish the asymptotic behavior of eigenvectors in a variety of situations.

We note that inconsistency of the sample eigenvectors does not necessarily imply poor performance of PCA. For example, PCA has been successfully applied in genome-wide association studies for accurate estimation of ethnicity (Price et al., 2006), and in PC regression for microarrays (Ma, Kosorok and Fine, 2006). However, for any individual study we cannot rule out the possibility of poor performance of the PC analysis. Our asymptotic result on the correlation coefficient between PC scores from sample and population eigenvectors provides us a measure to quantify the performance of PC analysis.

For the CEU/TSI data, SNP pruning was applied to adjust for strong LD among adjacent SNPs. Such SNP pruning is a common practice in the analysis of GWAS data, and has been implemented in the popular GWAS analysis software Plink (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly et al., 2007). The primary goal of SNP pruning is to avoid spurious PC results unrelated to population substructures. Technically, our approach does not rely on any independence assumption of the SNPs. However, strong local correlation may affect eigenvalues considerably. Thus the value in SNP pruning may be viewed as helping the data better accord with the assumptions of the spiked population model. From the CEU/TSI data and our experience in other GWAS data, we have found that the most common pruning

procedure implemented in Plink is sufficient for us to then apply our methods.

4.6 Proofs

Note that $\mathbf{E}\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Lambda}^{1/2}\mathbf{E}^T$ and $\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Lambda}^{1/2}$ have the same eigenvalues, and $\mathbf{E}^T\mathbf{U}$ is the eigenvector matrix of $\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Lambda}^{1/2}$. Since eigenvalues and angles between sample and population eigenvectors are what we concerned about, without loss of generality (WLOG), in the sequel, we assume $\mathbf{\Lambda}$ to be the population covariance matrix.

4.6.1 Notations

We largely follow notations in Paul (2007). We denote $\varphi_v(\mathbf{S})$ as the v^{th} largest eigenvalue of $\mathbf{S}$. Let suffice A represent the first m coordinates and B represent the remaining coordinates. Then we can partition $\mathbf{S}$ into

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{AA} & \mathbf{S}_{AB} \\ \mathbf{S}_{BA} & \mathbf{S}_{BB} \end{bmatrix}$$

We similarly partition the v^{th} eigenvector $\mathbf{u}_v^T$ into $(\mathbf{u}_{A,v}, \mathbf{u}_{B,v})$ and $\mathbf{Z}^T$ into $[\mathbf{Z}_A^T, \mathbf{Z}_B^T]$. Define R_v as $||\mathbf{u}_{B,v}||$ and let $\mathbf{a}_v = \mathbf{u}_{A,v}/\sqrt{1-R_v^2}$, then we get $||\mathbf{a}_v|| = 1$.

Applying singular value decomposition (SVD) to $\mathbf{Z}_B/\sqrt{n}$, we get

$$\frac{1}{\sqrt{n}}\mathbf{Z}_B = \mathbf{V}\mathbf{M}^{1/2}\mathbf{H}^T, \tag{4.9}$$

where $\mathbf{M} = diag(\mu_1, \ldots, \mu_{p-m})$ is a $(p-m) \times (p-m)$ diagonal matrix of ordered eigenvalues of $\mathbf{S}_{BB}$, $\mathbf{V}$ is a $(p-m)\times(p-m)$ orthogonal matrix, and $\mathbf{H}$ is an $n\times(p-m)$ matrix. For $n \geq p-m$, $\mathbf{H}$ has full rank orthogonal columns. When $n < p-m$, $\mathbf{H}$ has more columns than rows, hence it does not have full rank orthogonal columns. For the

later case, we make $\mathbf{H} = [\mathbf{H}_n, 0]$ where $\mathbf{H}_n$ is an $n \times n$ orthogonal matrix.

Propositions

We introduce two propositions for later use. The proofs of the 2 propositions can be found in section 4.6.3 and 4.6.3.

Proposition 1: Suppose $\mathbf{Y}$ is an $n \times m$ matrix with fixed m and each entry of $\mathbf{Y}$ is $i.i.d$ random variable which satisfies the moment condition of z_{ij} in Assumption 1. Let $\mathbf{C}$ be an $n \times n$ symmetric non-negative definite random matrix and independent of $\mathbf{Y}$. Further assume $||\mathbf{C}|| = O(1)$. Then

$$\frac{1}{n}\mathbf{Y}^T\mathbf{C}\mathbf{Y} - \frac{1}{n}trace(\mathbf{C})\mathbf{I} \xrightarrow{p} 0$$

as $n \to \infty$

Proposition 2: Suppose $\mathbf{y}$ is an n dimensional random vector which follows the same distribution of the row vectors of $\mathbf{Y}$ and independent of $\mathbf{S}_{BB}$. Let $f(x)$ be a bounded continuous function on $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$ and $f(0) = 0$. Suppose $\mathbf{F} = \text{diag}(f(\mu_1), \ldots, f(\mu_{p-m}))$, where $\{\mu_i\}_{i=1}^{p-m}$ are ordered eigenvalues of $\mathbf{M}$ which is defined on (4.9), then

$$\frac{1}{n}\mathbf{y}^T\mathbf{H}\mathbf{F}\mathbf{H}^T\mathbf{y} - \gamma \int f(x)dF_\gamma(x) \xrightarrow{p} 0$$

as $n \to \infty$, where $F_\gamma(x)$ is a distribution function of Marchenko-Pastur law with parameter γ (Marčenko and Pastur, 1967).

4.6.2 Proof of Part i) of Lemma 1

When p is fixed

By the strong law of large numbers, $\mathbf{S} \xrightarrow{a.s} \mathbf{\Lambda}$. Since eigenvalues are continuous with respect to the operator norm, the lemma follows after applying continuous mapping theorem.

When $p \to \infty$

For every small $\epsilon > 0$, there exist $\tilde{p}(n)$ and γ_ϵ such that $\tilde{p}(n)/n \to \gamma_\epsilon > 0$, $\lambda_v(1 + \gamma_\epsilon/(\lambda_v - 1)) < \lambda_v + \epsilon$ for all $v \leq m$, $(1 + \sqrt{\gamma_\epsilon})^2 < 1 + \epsilon$, and $(1 - \sqrt{\gamma_\epsilon})^2 > 1 - \epsilon$. For simplicity, we denote $\tilde{p}(n)$ as $\tilde{p}$. Suppose $\mathbf{Z}_{\tilde{p}}$ is a $\tilde{p} \times n$ matrix that satisfies the moment condition of z_{ij} in Assumption 1. Define an augmented data matrix $\tilde{\mathbf{X}}^T = [\mathbf{Z}^T\mathbf{\Lambda}, \mathbf{Z}_{\tilde{p}}^T]^T$ and its sample covariance matrix $\tilde{\mathbf{S}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$. Let $\mathbf{S}$ be a $p \times p$ upper left submatrix of $\tilde{\mathbf{S}}$. We also let $\hat{\mathbf{S}}$ be an $(m+1) \times (m+1)$ upper left submatrix of $\tilde{\mathbf{S}}$. For $v \leq (m+1)$, by the interlacing inequality (Theorem 4.3.15 of Horn and Johnson (1990)),

$$\varphi_v(\hat{\mathbf{S}}) \leq \varphi_v(\mathbf{S}) \leq \varphi_v(\tilde{\mathbf{S}}).$$

Since $\varphi_v(\hat{\mathbf{S}}) \xrightarrow{a.s} \lambda_v$, $\varphi_v(\tilde{\mathbf{S}}) \xrightarrow{a.s} \lambda_v(1 + \gamma_\epsilon/(\lambda_v - 1)) < 1 + \epsilon$ for $v \leq m$, and $\varphi_v(\tilde{\mathbf{S}}) \xrightarrow{a.s} (1 + \sqrt{\gamma_\epsilon})^2 < 1 + \epsilon$ for $v = m + 1$, we have

$$\lambda_v - o(1) \leq \lambda_v(\mathbf{S}) < \lambda_v + \epsilon + o(1), \text{ for } v \leq m + 1.$$

Thus,

$$\varphi_v(\mathbf{S}) \xrightarrow{a.s} \lambda_v, \text{ for } v \leq m + 1. \tag{4.10}$$

Similarly by the interlacing inequality, we get

$$\varphi_{\tilde{p}}(\tilde{S}) \leq \varphi_p(S) \leq \varphi_{m+1}(S).$$

Since $\varphi_{m+1}(S) \xrightarrow{a.s} 1$, and $\varphi_{\tilde{p}}(\tilde{S}) \xrightarrow{a.s} (1-\sqrt{\gamma_\epsilon})^2 > 1-\epsilon$, we conclude that

$$\varphi_p(S) \xrightarrow{a.s} 1. \tag{4.11}$$

The part i) of Lemma 1 follows by (4.10) and (4.11)

4.6.3 Proof of Part ii) of Lemma 2

Our proof of Lemma 2 (ii) closely follows the arguments in Paul (2007). From (Paul, 2007), it can be shown that

$$(\mathbf{S}_{AA} + \frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2})\mathbf{a}_v = d_v\mathbf{a}_v \tag{4.12}$$

and

$$\mathbf{a}_v^T(\mathbf{I} + \frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-2}\mathbf{H}^T\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2})\mathbf{a}_v = \frac{1}{1-R_v^2} \tag{4.13}$$

where $\mathbf{\Lambda}_A = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$.

When $\lambda_v > 1+\sqrt{\gamma}$

We can show that

$$\langle \mathbf{a}_v, \mathbf{e}_{A,v} \rangle \xrightarrow{p} 1 \tag{4.14}$$

and

$$\frac{1}{n}\mathbf{z}_{Av}^T\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-2}\mathbf{H}^T\mathbf{z}_{Av} \xrightarrow{p} \begin{cases} \gamma\int \frac{x}{(\rho_v-x)^2}dF_\gamma(x) & \text{for} \quad \gamma>0 \\ 0 & \text{for} \quad \gamma=0, \end{cases} \tag{4.15}$$

where $\mathbf{e}_{A,v}$ is a vector of the first m coordinates of the v^{th} population eigenvector $\mathbf{e}_v$, ρ_v is $\lambda_v\left(1+\frac{\gamma}{\lambda_v-1}\right)$, and $\mathbf{z}_{Av}$ is a vector of v^{th} row of $\mathbf{Z}_A$. The proofs can be found in 4.6.3. Note that $\mathbf{e}_v$ is a vector with 1 in its vth coordinate and 0 elsewhere. WLOG, we assume that $\langle\mathbf{e}_v,\mathbf{u}_v\rangle \geq 0$. Since $\langle\mathbf{e}_v,\mathbf{u}_v\rangle = \sqrt{1-R_v^2}\,\langle\mathbf{e}_{A,v},\mathbf{a}_v\rangle$, $\langle\mathbf{e}_v,\mathbf{u}_v\rangle \xrightarrow{p} \sqrt{1-R_v^2}$. By (4.13) and (4.15) , we can show that

$$\frac{1}{1-R_v^2} \xrightarrow{p} \begin{cases} 1+\lambda_v\gamma\int \frac{x}{(\rho_v-x)^2}dF_\gamma(x) & \text{for} \quad \gamma>0 \\ 1 & \text{for} \quad \gamma=0. \end{cases} \tag{4.16}$$

From Lemma B.2 of (Paul, 2007),

$$\int \frac{x}{(\rho_v-x)^2}dF_\gamma(x) = \frac{1}{(\lambda_v-1)^2-\gamma}. \tag{4.17}$$

Thus

$$\sqrt{1-R_v^2} \xrightarrow{p} \begin{cases} \sqrt{(1-\frac{\gamma}{(\lambda_v-1)^2})/(1+\frac{\gamma}{\lambda_v-1})} & \text{for} \quad \gamma>0 \\ 1 & \text{for} \quad \gamma=0. \end{cases} \tag{4.18}$$

It concludes the proof of the first part of Lemma 2 ii).

When $1<\lambda_v\leq 1+\sqrt{\gamma}$

Here we only need to consider $\gamma>0$ because no eigenvalue satisfies this condition when $\gamma=0$. We first show that $R_v \xrightarrow{p} 1$, which implies $\mathbf{u}_{A,v} \xrightarrow{p} 0$, hence $\langle\mathbf{e}_v,\mathbf{u}_v\rangle \xrightarrow{p} 0$. For

any $\epsilon > 0$ and $x \geq 0$, define

$$(x)_\epsilon = \begin{cases} x & \text{if } x > \epsilon \\ \epsilon & \text{if } x \leq \epsilon \end{cases}$$

and

$$\mathbf{G}_\epsilon = \text{diag}(d_v/((d_v - \mu_1)^2)_\epsilon, \ldots, d_v/((d_v - \mu_{p-m})^2)_\epsilon),$$

then by Propositions 1 and 2,

$$\frac{1}{n}\mathbf{z}_{Av}^T \mathbf{H}\mathbf{G}_\epsilon \mathbf{H}^T \mathbf{z}_{Av} \xrightarrow{p} \gamma \int \frac{x}{((\rho_v - x)^2)_\epsilon} dF_\gamma(x) \tag{4.19}$$

By monotone convergence theorem,

$$\gamma \int \frac{x}{((\rho_v - x)^2)_\epsilon} dF_\gamma(x) \xrightarrow{\epsilon \to 0} \gamma \int \frac{x}{(\rho_v - x)^2} dF_\gamma(x) \tag{4.20}$$

RHS of (4.20) is

$$\int_a^b \frac{\sqrt{(b-x)(x-a)}}{2\pi(\rho_v - x)^2} dx \tag{4.21}$$

where $a = (1 - \sqrt{\gamma})^2$ and $b = (1 + \sqrt{\gamma})^2$. Since (4.21) equals ∞ for any $a \leq \rho_v \leq b$, we conclude that

$$\frac{1}{n}\mathbf{z}_{Av}^T \mathbf{H}\mathbf{M}(d_v\mathbf{I} - \mathbf{M})^{-2}\mathbf{H}^T \mathbf{z}_{Av} \xrightarrow{p} \infty \tag{4.22}$$

Therefore, $R_v \xrightarrow{p} 1$, which proves the second part of Lemma 2 ii).

Proof of (4.14) **and** (4.15)

Define

$\mathcal{R}_v = \sum_{k \neq v}^{m} \frac{\lambda_v}{\rho_v(\lambda_k - \lambda_v)} \mathbf{e}_{A,k}\mathbf{e}_{A,k}^T$, $\mathcal{D}_v = \mathbf{S}_{AA} + \mathbf{S}_{AB}(d_v\mathbf{I} - \mathbf{S}_{BB})^{-1}\mathbf{S}_{BA} - (\rho_v/\lambda_v)\mathbf{\Lambda}_A$,
$\alpha_v = ||\mathcal{R}_v\mathcal{D}_v|| + |d_v - \rho_v|||\mathcal{R}_v||$, and $\beta_v = ||\mathcal{R}_v\mathcal{D}_v\mathbf{e}_{A,v}||$.

With the exactly same argument of (Paul, 2007), it can be shown that

$$\mathbf{a}_v - \mathbf{e}_{A,v} = -\mathcal{R}_v \mathcal{D}_v \mathbf{e}_{A,v} + \mathbf{r}_v$$

where $\mathbf{r}_v = -(1 - \langle \mathbf{e}_{A,v}, \mathbf{a}_v \rangle)\mathbf{e}_{A,v} - \mathcal{R}_v \mathcal{D}_v (\mathbf{a}_v - \mathbf{e}_{A,v}) + (d_v - \rho_v)\mathcal{R}_v(\mathbf{a}_v - \mathbf{e}_{A,v})$. By Lemma 1 of (Paul, 2005), $r_v = o_p(1)$, if $\alpha_v = o_p(1)$ and $\beta_v = o_p(1)$.

When $\gamma = 0$, $\mathbf{S}_{AA} - (\rho_v/\lambda_v)\mathbf{\Lambda}_A \xrightarrow{p} 0$ and the remainder of $\mathcal{D}_v$ is

$$\mathbf{S}_{AB}(d_v\mathbf{I} - \mathbf{S}_{BB})^{-1}\mathbf{S}_{BA} = \frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(d_v\mathbf{I} - \mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2}. \tag{4.23}$$

Since $d_v \xrightarrow{a.s} \lambda_v$ and $\mu_1 \xrightarrow{a.s} 1$,

$$||\mathbf{H}\mathbf{M}(d_v\mathbf{I} - \mathbf{M})^{-1}\mathbf{H}^T|| \xrightarrow{a.s} 1/(\lambda_v - 1)$$

By Proposition 1,

$$0 \leq ||(4.23)|| \leq \lambda_1 \frac{p\mu_1}{n(d_v - \mu_1)} + o_p(1) = o_p(1), \tag{4.24}$$

hence $\mathcal{D}_v = o_p(1)$.

When $\gamma > 0$, $\mathcal{D}_v$ can be written as

$$\begin{aligned}
\mathcal{D}_v &= [\mathbf{S}_{AA} - \mathbf{\Lambda}_A] \\
&+ [\mathbf{\Lambda}_A^{1/2}(\frac{1}{n}\mathbf{Z}_A\mathbf{H}\mathbf{M}(\rho_v\mathbf{I} - \mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A - \frac{1}{n}trace(\mathbf{M}(\rho_v\mathbf{I} - \mathbf{M})^{-1})\mathbf{I})\mathbf{\Lambda}_A^{1/2}] \\
&+ [(\frac{1}{n}trace(\mathbf{M}(\rho_v\mathbf{I} - \mathbf{M})^{-1}) - \gamma \int \frac{x}{\rho_v - x} dF_\gamma(x))\mathbf{\Lambda}_A] \\
&+ [(\rho_v - d_v)\frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(\rho_v\mathbf{I} - \mathbf{M})^{-1}(d_v\mathbf{I} - \mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A\mathbf{\Lambda}_A^{1/2}]
\end{aligned} \tag{4.25}$$

The first term of RHS is $o_p(1)$ by the weak law of large number. The second and third

terms are $o_p(1)$ by Propositions 1 and 2. For the fourth term, $\rho_v - d_v = o_p(1)$ and its remainder part is $O_p(1)$. Therefore, $\mathcal{D}_v = o_p(1)$. By combining the above results and $\mathcal{R}_v = O_p(1)$ plus $d_v - \rho_v = o_p(1)$, we prove the Equation (4.14).

For (4.15): When $\gamma = 0$, (4.15) can be proved by the exactly same way used to show (4.24). When $\gamma > 0$, $d_v \xrightarrow{a.s} \rho_v$, and $\mu_1 \xrightarrow{a.s} (1+\sqrt{\gamma})^2 < \rho_v$, hence $||\mathbf{C}|| \xrightarrow{a.s} \frac{(1+\sqrt{\gamma})^2}{(\rho_v-(1+\sqrt{\gamma})^2)^2}$. Therefore, the result follows according to Propositions 1 and 2.

Proof of Proposition 1

Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ be the ordered eigenvalues of $\mathbf{C}$, and c_{ij} be the (i,j)th element of $\mathbf{C}$. Suppose $\mathbf{y}_s$ is the sth column of $\mathbf{Y}$, and y_{ij} is the (i,j)th element of $\mathbf{Y}$. We further define $\psi(s,s) = \frac{1}{n}\mathbf{y}_s^T\mathbf{C}\mathbf{y}_s - \frac{1}{n}trace(\mathbf{C})$ and $\psi(s,t) = \frac{1}{n}\mathbf{y}_s^T\mathbf{C}\mathbf{y}_t$ for $s \neq t$. The conditional mean of $\psi(s,s)$ given $\mathbf{C}$ is

$$\begin{aligned} E(\psi(s,s)|\mathbf{C}) &= E(\frac{1}{n}\sum_{i,j} c_{ij}y_{is}y_{js}|\mathbf{C}) - \frac{1}{n}\sum_{i=1}^{n}\mu_i \\ &= \frac{1}{n}\sum_{i=1}^{n} c_{ii}E(y_{is}^2) + \frac{2}{n}\sum_{i<j}^{n} c_{ij}E(y_{is}y_{js}) - \frac{1}{n}\sum_{i=1}^{n}\mu_i \\ &= \frac{1}{n}\sum_{i=1}^{n} c_{ii} - \frac{1}{n}\sum_{i=1}^{n}\mu_i = 0 \end{aligned} \tag{4.26}$$

Thus, $E(\psi(s,s)) = E(E(\psi(s,s)|\mathbf{C})) = E(0) = 0$.

Next, the conditional variance of $\psi(s,s)$ given $\mathbf{C}$ is

$$\begin{aligned} Var(\psi(s,s)|\mathbf{C}) &= \frac{1}{n^2} Var(\sum_{i,j} c_{ij} y_{is} y_{js} | \mathbf{C}) \\ &= \frac{1}{n^2} \sum_{i,j,l,q=1}^{n} c_{ij} c_{lq} Cov(y_{is} y_{js}, y_{ls} y_{qs}) \\ &= \frac{4}{n^2} \sum_{i,j=1}^{n} c_{ij}^2 Var(y_{is} y_{js}) \\ &\leq \frac{4\alpha}{n^2} \sum_{i,j=1}^{n} c_{ij}^2 = \frac{4\alpha}{n^2} trace(\mathbf{C}^2) = \frac{4\alpha}{n^2} \sum_{i=1}^{n} \mu_i^2 \end{aligned} \tag{4.27}$$

where $\alpha = max(1, E(y_{is}^4) - 1)$. Since $||\mathbf{C}|| = O(1)$, $\mu_i^2 \leq ||\mathbf{C}||^2 = O(1)$. Therefore, $Var(\psi(s,s)|\mathbf{C}) \leq O(1/n)$ and $Var(\psi(s,s)) = Var(E(\psi(s,s)|\mathbf{C})) + E(Var(\psi(s,s)|\mathbf{C})) \leq 0 + O(1/n) \to 0$ as $n \to \infty$. By Chebyshev inequality, we can conclude that

$$\psi(s,s) \xrightarrow{p} 0.$$

We can similarly show $\psi(s,t) \xrightarrow{p} 0$, which we omit here.

Proof of Proposition 2

Consider an expansion

$$\begin{aligned} &\frac{1}{n} y^T \mathbf{H}\mathbf{F}\mathbf{H}^T y - \gamma \int f(x) dF_\gamma(x) \\ &= [\frac{1}{n} y^T \mathbf{H}\mathbf{F}\mathbf{H}^T y - \frac{1}{n} trace(\mathbf{F})] \\ &+ [\frac{1}{n} trace(\mathbf{F}) - \gamma \int f(x) dF_\gamma(x)] \\ &= (a) + (b) \end{aligned}$$

We show that both (a) and (b) converge to 0 in probability.

(a) : Since $\mu_1 \xrightarrow{a.s.} (1+\sqrt{\gamma})^2$, $\mu_{min(p-m,n)} \xrightarrow{a.s.} (1-\sqrt{\gamma})^2$, $\mu_k = 0$ for $k > min(p-m, n)$,

and $f(x)$ is continuous and bounded on $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$, there exists $K > 0$ such that $\sup_i |f(\mu_i)| < K$ a.s. Let $\mathbf{C} = \mathbf{HFH}^T$, then $trace(\mathbf{C}) = trace(\mathbf{F})$. By Proposition 1, $(a) = o_p(1)$.

(b) : Let F_{p-m} be an empirical spectral distribution of $\mathbf{S}_{BB}$, then

$$\frac{1}{n} trace(\mathbf{F}) = \frac{p-m}{n} \int f(x) dF_{p-m}(x),$$

and $\int f(x) dF_n(x) \xrightarrow{p} \int f(x) dF_\gamma(x)$ (Marčenko and Pastur, 1967; Bai, 1999). Thus

$$\frac{p-m}{n} \int f(x) dF_{p-m}(x) \xrightarrow{p} \gamma \int f(x) dF_\gamma(x),$$

which shows that $(b) = o_p(1)$.

Combining (a) and (b), we finish the proof.

4.6.4 Proof of Theorem 2

WLOG we assume $\langle \mathbf{g}_v, \tilde{\mathbf{p}}_v \rangle \geq 0$. Let $\mathbf{e}_v = \{\mathbf{e}_{A,v}, \mathbf{e}_{B,v}\}$, then $\mathbf{e}_{A,v}$ is the vector with 1 in v^{th} coordinate and 0 elsewhere, and $\mathbf{e}_{B,v}$ is the zero vector. Since $\mathbf{S}_{AA}\mathbf{u}_{A,v} + \mathbf{S}_{AB}\mathbf{u}_{B,v} = d_v \mathbf{u}_{A,v}$, we have

$$\begin{aligned}
< \mathbf{g}_v, \tilde{\mathbf{p}}_v > &= \frac{1}{n\sqrt{d_v \lambda_v}} \mathbf{e}_v^T \mathbf{X}\mathbf{X}^T \mathbf{u}_v \\
&= \mathbf{e}_{A,v}^T \mathbf{S}_{AA} \mathbf{u}_{A,v} / \sqrt{d_v \lambda_v} + \mathbf{e}_{A,v}^T \mathbf{S}_{AB} \mathbf{u}_{B,v} / \sqrt{d_v \lambda_v} \\
&= \frac{d_v}{\sqrt{d_v \lambda_v}} \mathbf{e}_{A,v}^T \mathbf{u}_{A,v} = \sqrt{\frac{d_v}{\lambda_v}} \mathbf{e}_v^T \mathbf{u}_v \\
&\xrightarrow{p} \begin{cases} \sqrt{(1 - \frac{\gamma}{(\lambda_v - 1)^2})} & \text{for} \quad \lambda_v > 1 + \sqrt{\gamma} \\ 0 & \text{for} \quad 1 < \lambda_v \leq 1 + \sqrt{\gamma}. \end{cases}
\end{aligned} \tag{4.28}$$

4.6.5 Proof of Theorem 3

First, we show the square of the denominator converges to $\rho(\lambda_v)$. Since $p_{vj} = \mathbf{u}_v^T \mathbf{x}_j$, and $E(p_{vi}^2) = E(p_{vj}^2)$ for $i \neq j$,

$$\begin{aligned} E(p_{vj}^2) &= \frac{1}{n} E(\sum_{j=1}^{n} p_{vj}^2) = \frac{1}{n} E(\sum_{j=1}^{n} (\mathbf{u}_v^T \mathbf{x}_j)^2) \\ &= E(\mathbf{u}_v^T \mathbf{X}\mathbf{X}^T \mathbf{u}_v / n) = E(d_v) \xrightarrow{a.s} \rho(\lambda_v) \end{aligned} \tag{4.29}$$

Next we show the square of numerator converges to $\phi(\lambda_v)^2(\lambda_v - 1) + 1$. Define $\mathbf{u}_v^{\perp} := \frac{1}{\sqrt{1-(\mathbf{u}_v^T \mathbf{e}_v)^2}}(I - \mathbf{e}_v \mathbf{e}_v^T)\mathbf{u}_v$, then $\mathbf{u}_v$ can be expressed as

$$\mathbf{u}_v = (\mathbf{u}_v^T \mathbf{e}_v)\mathbf{e}_v + \sqrt{1 - (\mathbf{u}_v^T \mathbf{e}_v)^2}\mathbf{u}_v^{\perp}$$

Partition $\mathbf{u}_v^{\perp} = \left\{\mathbf{u}_{A,v}^{\perp}, \mathbf{u}_{B,v}^{\perp}\right\}$. From (4.14), $\mathbf{a}_v \xrightarrow{p} \mathbf{e}_{A,v}$, therefore $\mathbf{u}_{A,v}^{\perp} \xrightarrow{p} 0$ and $\mathbf{u}_{B,v}^{\perp T}\mathbf{u}_{B,v}^{\perp} \xrightarrow{p} 1$. Since $\mathbf{x}_{new}$ and $\mathbf{u}_v$ are independent, we have

$$\begin{aligned} E(q_v^2|\mathbf{u}_v) &= E((\mathbf{u}_v^T \mathbf{x}_{new})^2|\mathbf{u}_v) = \mathbf{u}_v^T E(\mathbf{x}_{new}\mathbf{x}_{new}^T|\mathbf{u}_v)\mathbf{u}_v = \mathbf{u}_v^T \Lambda \mathbf{u}_v \\ &= (\mathbf{u}_v^T \mathbf{e}_v)^2 \mathbf{e}_v^T \Lambda \mathbf{e}_v + (1 - (\mathbf{u}_v^T \mathbf{e}_v)^2)\mathbf{u}_v^{\perp T}\Lambda \mathbf{u}_v^{\perp} + 2\mathbf{u}_v^T \mathbf{e}_v \sqrt{1 - (\mathbf{u}_v^T \mathbf{e}_v)^2}\mathbf{e}_v^T \Lambda \mathbf{u}_v^{\perp} \\ &= (\mathbf{u}_v^T \mathbf{e}_v)^2 \lambda_v + (1 - (\mathbf{u}_v^T \mathbf{e}_v)^2)(\mathbf{u}_{A,v}^{\perp T}\Lambda_A \mathbf{u}_{A,v}^{\perp} + \mathbf{u}_{B,v}^{\perp T}\mathbf{u}_{B,v}^{\perp}) \\ &+ 2\mathbf{u}_v^T \mathbf{e}_v \sqrt{1 - (\mathbf{u}_v^T \mathbf{e}_v)^2}\mathbf{e}_{A,v}\Lambda_A \mathbf{u}_{A,v}^{\perp} \\ &\xrightarrow{p} \phi(\lambda_v)^2(\lambda_v - 1) + 1. \end{aligned} \tag{4.30}$$

From (4.29) and (4.30),

$$\sqrt{\frac{E(q_v^2)}{E(p_{vi}^2)}} \to \sqrt{\frac{\phi(\lambda_v)^2(\lambda_v - 1) + 1}{\rho(\lambda_v)}} = \frac{(\lambda_v - 1)}{(\lambda_v + \gamma - 1)}. \tag{4.31}$$

4.6.6 Proof of Theorem 4

Since $\rho^{-1}(pr_v) \to \lambda_v$ for $v \leq k$, WLOG we assume that $k_0 = k$, where k is the number of λ_v bigger than $1 + \sqrt{\gamma}$. Set

$$h(x) = \sum_{v=1}^{k} \rho^{-1}(r_v x) + p - k - x \tag{4.32}$$

The first and second partial derivatives of $h(x)$ are

$$\frac{\partial h(x)}{\partial x} = \frac{1}{2}\sum_{v=1}^{k} r_v + \frac{1}{2}\sum_{v=1}^{k} \frac{(xr_v - (1+\gamma))r_v}{\sqrt{(xr_v - (1+\gamma))^2 - 4\gamma}} - 1 \tag{4.33}$$

$$\frac{\partial^2 h(x)}{\partial x^2} = 2\sum_{v=1}^{k} \frac{-r_v^2\gamma}{((xr_v - (1+\gamma))^2 - 4\gamma)^{3/2}} < 0, \tag{4.34}$$

so $h(x)$ is a concave function of x given r_v. From the fact that $\rho^{-1}(r_v p) > 1$ for $v \leq k$, we know $h(p) > 0$. Because of the concave nature of this function, $h(x) = 0$ has an unique solution τ on $[p, \infty)$, which $\sum_{v=1}^{k_l} \hat{\lambda}_{v,l} + p - m_l$ converges to. Thus $\hat{d}_v = \tau r_v$. Define $\tilde{d}_v = r_v\omega$ where $\omega = \sum_{v=1}^{k} \lambda_v + p - k$, and set d_v as the sample eigenvalue when $\sigma^2 = 1$. The sum of all d_v is

$$\sum_{v=1}^{p} d_v = \frac{1}{n} trace(ZZ^T\Lambda) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n} \lambda_i z_{ij}^2, \tag{4.35}$$

thus

$$E\left(\frac{\sum_{v=1}^{p} d_v}{\omega}\right) = \frac{\sum_{v=1}^{m} \lambda_v + p - k}{\omega} \to 1, \text{ and} \tag{4.36}$$

$$Var\left(\frac{\sum_{v=1}^{p} d_v}{\omega}\right) = \frac{1}{n}\frac{\sum_{v=1}^{p} \lambda_v^2}{\omega^2}(E(z_{11}^4) - 1) \to 0. \tag{4.37}$$

By (4.36) and (4.37)

$$\sum_{v=1}^{p} d_v/\omega = 1 + o_p(1). \tag{4.38}$$

Since $d_v \to \rho(\lambda_v)$ for $v \leq k$,

$$\tilde{d}_v = d_v\omega / \sum_{v=1}^{p} d_v = d_v(1 + o_p(1)) \xrightarrow{p} \rho(\lambda_v). \tag{4.39}$$

Now, we show that $\tau = \omega + o_p(1)$. Plugging ω into $h(x)$ and combining the fact that $\rho^{-1}(\tilde{d}_v) = \lambda_v + o_p(1)$, we get

$$h(\omega) = \sum_{v=1}^{k} \rho^{-1}(\tilde{d}_v) - \sum_{v=1}^{k} \lambda_v = o_p(1). \tag{4.40}$$

From the facts that $h(x)$ is a continuous concave function, $\omega > p$, and $h(p) > 0$, we conclude that

$$\omega = \tau + o_p(1). \tag{4.41}$$

Therefore,

$$\hat{d}_v = r_v\tau = r_v(\omega + op(1)) = \tilde{d}_v + op(1) \xrightarrow{p} \rho(\lambda_v) \tag{4.42}$$

for $v \leq k$, which concludes the proof.

4.6.7 Proof of Theorem 5

The v^{th} eigenvalue of the sample covariance matrix S is

$$\begin{aligned} d_v = \varphi_v(\mathbf{S}) &= \varphi_v(\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{\Lambda}^{1/2}/n) = \varphi_v(\mathbf{Z}^T\mathbf{\Lambda}\mathbf{Z}/n) \\ &= \varphi_v(\mathbf{Z}_A^T\mathbf{\Lambda}_A\mathbf{Z}_A/n + \mathbf{Z}_B^T\mathbf{Z}_B/n) \end{aligned}$$

From Proposition 1,

$$\varphi_v(\mathbf{Z}_B^T\mathbf{Z}_B/n)/\hat{\gamma} - 1 = o(1), \tag{4.43}$$

for all $v = 1, \ldots, n$.

First, we consider $c_v \gg o(1)$. By the strong law of large number,

$$\mathbf{Z}_A\mathbf{Z}_A^T/n - \mathbf{I}_{m\times m} = o(1),$$

and by the same increment rate of spiked population eigenvalues,

$$||\mathbf{\Lambda}_A/\lambda_v|| = O(1),$$

for $v \leq m$. Therefore,

$$\begin{aligned}\left\|\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2}/n\lambda_v - \mathbf{\Lambda}_A/\lambda_v\right\| &= \left\|\mathbf{\Lambda}_A^{1/2}(\mathbf{Z}_A\mathbf{Z}_A^T/n - \mathbf{I})\mathbf{\Lambda}_A^{1/2}/\lambda_v\right\| \\ &= o(1)\end{aligned} \tag{4.44}$$

By the continuity of eigenvalues,

$$\begin{aligned}\frac{\varphi_v(\mathbf{Z}_A^T\mathbf{\Lambda}_A\mathbf{Z}_A/n)}{\lambda_v} - 1 = \frac{\varphi_v(\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2}/n)}{\lambda_v} - 1 = 1 + o(1) - 1 \\ = o(1)\end{aligned} \tag{4.45}$$

for $v \leq m$, and

$$\lambda_v(\mathbf{Z}_A^T\mathbf{\Lambda}_A\mathbf{Z}_A/n) = 0, \tag{4.46}$$

for $v > m$, because the rank of $\mathbf{Z}_A^T\mathbf{\Lambda}_A\mathbf{Z}_A/n$ is m. By Weyl's inequality (Bhatia, 1997) $\varphi_v(\mathbf{S}) \leq \varphi_v(\mathbf{Z}_A^T\mathbf{\Lambda}_A\mathbf{Z}_A/n) + \varphi_1(\mathbf{Z}_B^T\mathbf{Z}_B/n)$ and $\varphi_v(\mathbf{S}) \geq \varphi_v(\mathbf{Z}_A^T\mathbf{\Lambda}_A\mathbf{Z}_A/n) + \varphi_1(\mathbf{Z}_B^T\mathbf{Z}_B/n)$.

Thus,

$$\frac{d_v}{\lambda_v} - \frac{c_v + 1}{c_v} = o(1)$$

for $v \leq m$, and

$$\frac{d_v}{\hat{\gamma}} \xrightarrow{a.s} 1$$

for $v > m$. Now, we consider $c_v = o(1)$. It can be easily shown

$$\frac{\varphi_1(\mathbf{Z}_A^T \mathbf{\Lambda}_A \mathbf{Z}_A / n)}{\hat{\gamma}} \xrightarrow{a.s} 0.$$

Thus,

$$\frac{d_v}{\hat{\gamma}} \xrightarrow{a.s} 1,$$

which concludes the proof.

4.6.8 Proof of Theorem 6

1) When c_v is bounded away from zero and $v \leq m$: Define

$$\eta_v = \frac{c_v + 1}{c_v} \tag{4.47}$$

$$\mathcal{R}_v^* = \sum_{k \neq v}^{m} \frac{\lambda_v}{\eta_v(\lambda_k - \lambda_v)} \mathbf{e}_{A,k} \mathbf{e}_{A,k}^T, \tag{4.48}$$

and

$$\mathcal{D}_v^* = \frac{1}{\lambda_v}(\mathbf{S}_{AA} + \mathbf{S}_{AB}(d_v \mathbf{I} - \mathbf{S}_{BB})^{-1} \mathbf{S}_{BA} - \eta_v \mathbf{\Lambda}_A), \tag{4.49}$$

where $\mathbf{e}_{A,k}$ is the first m elements of $\mathbf{e}_v$. From (4.12),

$$\left(\frac{\eta_v}{\lambda_v} \mathbf{\Lambda}_{AA} - \eta_v \mathbf{I}\right) = -\mathcal{D}_v^* \mathbf{a}_v + \left(\frac{d_v}{\lambda_v} - \eta_v\right) \mathbf{a}_v \tag{4.50}$$

Since

$$\mathcal{R}_v^*\left(\frac{\eta_v}{\lambda_v}\mathbf{\Lambda}_{AA}-\eta_v\mathbf{I}\right)=\sum_{k\neq v}^{m}\left(\frac{\lambda_v}{\eta_v(\lambda_k-\lambda_v)}\mathbf{e}_{A,k}\mathbf{e}_{A,k}^T-\frac{\lambda_v}{\lambda_k-\lambda_v}\mathbf{e}_{A,k}\mathbf{e}_{A,k}^T\right)$$
$$=\sum_{k\neq v}^{m}\mathbf{e}_{A,k}\mathbf{e}_{A,k}^T, \tag{4.51}$$

it can be shown

$$(\mathbf{I}-\mathbf{e}_{A,v}\mathbf{e}_{A,v}^T)\mathbf{a}_v=-\mathcal{R}_v^*\mathcal{D}_v^*\mathbf{a}_v+\left(\frac{d_v}{\lambda_v}-\eta_v\right)\mathcal{R}_v^*\mathbf{a}_v \tag{4.52}$$

(4.52) indicates $a_v-e_v=op(1)$, if both $||\mathcal{R}_v^*\mathcal{D}_v^*||$ and $|d_v/\lambda_v-\eta_v|||\mathcal{R}_v^*||$ are $op(1)$. For $l=1,\ldots,m$,

$$\mathcal{R}_v^*\mathcal{D}_v^*\mathbf{e}_{A,l}=\sum_{k\neq v}^{m}\frac{\lambda_v}{\eta_v(\lambda_k-\lambda_v)}\mathbf{e}_{A,k}^T\mathcal{D}_v^*\mathbf{e}_{A,l} \tag{4.53}$$

and

$$\begin{aligned}\mathbf{e}_{A,k}^T\mathcal{D}_v^*\mathbf{e}_{A,l}&=\mathbf{e}_{A,k}^T\mathbf{S}_{AA}\mathbf{e}_{A,l}/\lambda_v\\&+\mathbf{e}_{A,k}^T\mathbf{S}_{AB}(d_v\mathbf{I}-\mathbf{S}_{BB})^{-1}S_{BA}\mathbf{e}_{A,l}/\lambda_v\\&-\eta_v\mathbf{e}_{A,k}^T\mathbf{\Lambda}_A\mathbf{e}_{A,l}/\lambda_v\end{aligned} \tag{4.54}$$

The first term is

$$\frac{1}{\lambda_v}\mathbf{e}_{A,k}^T\mathbf{S}_{AA}\mathbf{e}_{A,l}=\frac{\lambda_k^{1/2}\lambda_l^{1/2}}{\lambda_v}\frac{1}{n}\mathbf{z}_{A,k}^T\mathbf{z}_{A,l}=\begin{cases}\frac{\lambda_k^{1/2}\lambda_l^{1/2}}{\lambda_v}op(1) & ,\text{if } k\neq l\\ \frac{\lambda_k}{\lambda_v}(1+op(1)) & ,\text{if } k=l,\end{cases}. \tag{4.55}$$

The second term is

$$\frac{1}{\lambda_v}\mathbf{e}_{A,k}^T\mathbf{S}_{AB}(d_v\mathbf{I}-\mathbf{S}_{BB})^{-1}S_{BA}\mathbf{e}_{A,v}=\frac{\lambda_k^{1/2}\lambda_l^{1/2}}{\lambda_v}\frac{1}{n}\mathbf{z}_{A,k}^T\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{z}_{A,l} \quad (4.56)$$

From the Proposition 1,

$$\frac{1}{n}\mathbf{z}_{A,k}^T\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{z}_{A,l}=op(1)$$

for $l\neq k$, and

$$\frac{1}{n}\mathbf{z}_{A,k}^T\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{z}_{A,l}=\frac{1}{n}trace(\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1})+op(1)$$

for $l=k$. Since $trace(\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1})/n<\mu_1/(d_v-\mu_1)$, $trace(\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1})/n>\mu_n/(d_v-\mu_n)$, and both $\mu_1/\hat{\gamma}$ and $\mu_n/\hat{\gamma}$ are $o(1)$,

$$\begin{aligned}\frac{1}{n}trace(\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1})&=\frac{\mu_1/\hat{\gamma}}{d_v/\hat{\gamma}-\mu_1/\hat{\gamma}}+o(1)\\&=\frac{1+o(1)}{c_v+c_vo(1)}+o(1).\end{aligned} \quad (4.57)$$

Hence,

$$(4.56)=\begin{cases}\frac{\lambda_k^{1/2}\lambda_l^{1/2}}{\lambda_v}op(1) & ,\text{if } k\neq l\\ \frac{\lambda_k}{\lambda_v}\left(\frac{1+o(1)}{c_v+c_vo(1)}+op(1)\right) & ,\text{if } k=l\end{cases}. \quad (4.58)$$

The third term is

$$\frac{\eta_v}{\lambda_v}\mathbf{e}_{A,k}^T\mathbf{\Lambda}_A\mathbf{e}_{A,l}=\begin{cases}0 & ,\text{if } k\neq l\\ \frac{\eta_v}{\lambda_v}\lambda_k & ,\text{if } k=l,\end{cases}. \quad (4.59)$$

Combining (4.55), (4.56), and (4.59), it can be shown

$$\mathbf{e}_{A,k}^T\mathcal{D}_v^*\mathbf{e}_{A,l}=\frac{\lambda_k^{1/2}\lambda_l^{1/2}}{\lambda_v}op(1) \quad (4.60)$$

for $l, k = 1, \ldots, m$. Thus, $||\mathcal{R}_v^* \mathcal{D}_v^*|| = op(1)$. Since $||\mathcal{R}_v^*|| = O(1)$,

$$\left| \frac{d_v}{\lambda_v} - \eta_v \right| ||\mathcal{R}_v^*|| = op(1) \tag{4.61}$$

Hence,

$$a_v - e_v = op(1) \tag{4.62}$$

By the exactly same logic used in (4.57),

$$\frac{\lambda_v}{n} \mathbf{z}_{Av}^T \mathbf{H}\mathbf{M}(d_v \mathbf{I} - \mathbf{M})^{-2} \mathbf{H}^T \mathbf{z}_{Av} = \frac{1 + o(1)}{c_v + c_v o(1)} + op(1) = \frac{1}{c_v} + op(1) \tag{4.63}$$

From (4.62)

$$\begin{aligned} (4.15) &= (e_v + op(1))^T (\mathbf{I} + \frac{1}{n} \mathbf{\Lambda}_A^{1/2} \mathbf{Z}_A \mathbf{H}\mathbf{M}(d_v \mathbf{I} - \mathbf{M})^{-2} \mathbf{H}^T \mathbf{Z}_A^T \mathbf{\Lambda}_A^{1/2})(e_v + op(1)) \\ &= 1 + \frac{1}{c_v} + op(1). \end{aligned} \tag{4.64}$$

The result follows by combining (4.62) and (4.64).

2) When $c_v = o(1)$: Since $\mathbf{e}_1^T \mathbf{u}_1 < \sqrt{1 - R_1}$, we only need to show $R_1 \to 1$. By the definition of sample eigenvalues and eigenvectors,

$$\begin{aligned} d_1 &= \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 \\ &= \mathbf{u}_{A,1}^T \mathbf{S}_{AA} \mathbf{u}_{A,1} + 2\mathbf{u}_{A,1}^T \mathbf{S}_{AB} \mathbf{u}_{B,1} + \mathbf{u}_{B,1}^T \mathbf{S}_{BB} \mathbf{u}_{B,1} \end{aligned} \tag{4.65}$$

The first term of (4.65) is

$$\begin{aligned} \mathbf{u}_{A,1}^T \mathbf{S}_{AA} \mathbf{u}_{A,1} &= \mathbf{u}_{A,1}^T \mathbf{\Lambda}_{AA}^{1/2} \mathbf{Z}_A \mathbf{Z}_A^T \mathbf{\Lambda}_{AA}^{1/2} \mathbf{u}_{A,1} / n \\ &\leq (1 - R_1^2) \lambda_1 (\mathbf{\Lambda}_{AA}^{1/2} \mathbf{Z}_A \mathbf{Z}_A^T \mathbf{\Lambda}_{AA}^{1/2} / n) \qquad \leq (1 - R_1^2) \lambda_1 (1 + o(1)) \end{aligned} \tag{4.66}$$

The second term of (4.65) is

$$\mathbf{u}_{A,1}^T \mathbf{S}_{AB} \mathbf{u}_{B,1} = \mathbf{u}_{A,1}^T \mathbf{\Lambda}_{AA}^{1/2} \mathbf{Z}_A \mathbf{Z}_B^T \mathbf{u}_{B,1}/n \tag{4.67}$$

Since $\mathbf{u}_{B,1}/R_1$ is an unit vector, $\mathbf{z}_v^T \mathbf{Z}_B^T \mathbf{u}_{B,1}/n \leq R_1 \sqrt{\mathbf{z}_v^T \mathbf{Z}_B^T \mathbf{Z}_B \mathbf{z}_v/n^2}$, where $\mathbf{z}_v$ is the v^{th} row of $\mathbf{Z}$. Hence,

$$\begin{aligned}(4.67) &\leq R_1 \sqrt{\lambda_1 (1 - R_1^2) \sum_{v=1}^{m} \mathbf{z}_v^T \mathbf{Z}_B^T \mathbf{Z}_B \mathbf{z}_v/n^2} = R_1 \sqrt{\lambda_1 \hat{\gamma} (1 - R_1^2) \sum_{v=1}^{m} \mathbf{z}_v^T \mathbf{Z}_B^T \mathbf{Z}_B \mathbf{z}_v/np} \\ &= R_1 \sqrt{\lambda_1 \hat{\gamma} (1 - R_1^2)} Op(1) \end{aligned} \tag{4.68}$$

The third term of (4.65) is

$$\mathbf{u}_{B,1}^T \mathbf{S}_{BB} \mathbf{u}_{B,1} < R_1^2 \mu_1 \tag{4.69}$$

From (4.66), (4.68) and (4.69),

$$\frac{d_1 - \mu_1}{\hat{\gamma}} \leq (1 - R_1^2) c_1 (1 + o(1)) + R_1 \sqrt{1 - R_1^2} \sqrt{c_1} Op(1) + (R_1^2 - 1)(1 + o(1))$$

By the interlacing inequality (Theorem 4.3.15 of Horn and Johnson (1990)), $d_1 - \mu_1 \geq 0$. Combining the fact that $c_v = o(1)$, we can conclude that $R_1 \to 1$, which complete the proof.

4.6.9 Proof of Theorem 7

WLOG we assume $\langle \mathbf{g}_v, \tilde{\mathbf{p}}_v \rangle \geq 0$. Since $\mathbf{S}_{AA}\mathbf{u}_{A,v} + \mathbf{S}_{AB}\mathbf{u}_{B,v} = d_v \mathbf{u}_{A,v}$,

$$\begin{aligned}
\langle \mathbf{g}_v, \tilde{\mathbf{p}}_v \rangle &= \frac{\mathbf{g}_v^T \tilde{\mathbf{p}}_v}{\sqrt{\mathbf{g}_v^T \mathbf{g}_v}} = \frac{1}{n\sqrt{d_v \lambda_v}} \mathbf{e}_v^T \mathbf{X}\mathbf{X}^T \mathbf{u}_v \\
&= \frac{1}{n\sqrt{d_v \lambda_v}} \mathbf{e}_{A,v}^T \mathbf{S}_{AA}\mathbf{u}_{A,v} + \frac{1}{n\sqrt{d_v \lambda_v}} \mathbf{e}_{A,v}^T \mathbf{S}_{AB}\mathbf{u}_{B,v} \\
&= \frac{d_v}{n\sqrt{d_v \lambda_v}} \mathbf{e}_{A,v}^T \mathbf{u}_{A,v} \\
&= \sqrt{\frac{d_v}{\lambda_v}} \mathbf{e}_v^T \mathbf{u}_v
\end{aligned} \tag{4.70}$$

Since $\mathbf{e}_v^T \mathbf{u}_v \xrightarrow{p} \sqrt{c_v/(c_v+1)}$ and $d_v/\lambda_v \to (c_v+1)/c_v$, (4.70) converges to 1 in probability, which completes the proof.

4.6.10 Proof of Theorem 8

Using the exactly same argument in section 4.6.5, it can be shown that

$$\frac{E(p_{vj}^2)}{\lambda_v} = \frac{E(d_v)}{\lambda_v} \to \frac{c_v + 1}{c_v} \tag{4.71}$$

and

$$\begin{aligned}
\frac{E(q_v^2 | \mathbf{u}_v)}{\lambda_v} &= (\mathbf{u}_v^T \mathbf{e}_v)^2 + \frac{1}{\lambda_v}(1 - (\mathbf{u}_v^T \mathbf{e}_v)^2)(\mathbf{u}_{A,v}^{\perp T} \Lambda_A \mathbf{u}_{A,v}^{\perp} + \mathbf{u}_{B,v}^{\perp T} \mathbf{u}_{B,v}^{\perp}) \\
&\quad + \frac{2}{\lambda_v} \mathbf{u}_v^T \mathbf{e}_v \sqrt{1 - (\mathbf{u}_v^T \mathbf{e}_v)^2} \mathbf{e}_{A,v} \Lambda_A \mathbf{u}_{A,v}^{\perp} \\
&\xrightarrow{p} \frac{c_v}{1 + c_v}
\end{aligned} \tag{4.72}$$

From (4.71) and (4.72)

$$\sqrt{\frac{E(q_v^2)}{E(p_{vi}^2)}} \to \frac{c_v}{1 + c_v} \tag{4.73}$$

which concludes the proof.

4.7 Tables and Figures

Table 4.1: Cosine angle estimates of eigenvectors and PC scores based on 1000 simulations. "Angle" indicates the theoretical asymptotic $cos(angle)$, "Estimate1" indicates the empirical $cos(angle)$ estimator, 'Estimate2" indicates the asymptotic $cos(angle)$ estimator. For each estimator, each entry represents mean of $1,000$ simulation results with standard error in parentheses.

		PC 1			PC 2		
γ	n	Angle	Angle Estimate1	Angle Estimate2	Angle	Angle Estimate1	Angle Estimate2
Eigenvectors							
1	100	0.93	0.93(0.013)	0.91(0.027)	0.82	0.81(0.053)	0.80(0.052)
	200		0.93(0.009)	0.92(0.014)		0.81(0.030)	0.81(0.032)
20	100	0.70	0.69(0.037)	0.70(0.031)	0.51	0.50(0.053)	0.50(0.058)
	200		0.69(0.023)	0.70(0.022)		0.51(0.036)	0.51(0.041)
100	100	0.53	0.53(0.034)	0.53(0.031)	0.37	0.35(0.043)	0.35(0.047)
	200		0.53(0.024)	0.53(0.024)		0.36(0.029)	0.36(0.033)
500	100	0.38	0.38(0.029)	0.38(0.028)	0.25	0.24(0.033)	0.24(0.037)
	200		0.38(0.020)	0.38(0.020)		0.25(0.021)	0.25(0.024)
PC Scores							
1	100	0.99	0.99(0.004)	0.98(0.016)	0.94	0.93(0.036)	0.91(0.048)
	200		0.99(0.003)	0.99(0.006)		0.94(0.019)	0.93(0.024)
20	100	0.98	0.97(0.083)	0.98(0.008)	0.89	0.86(0.105)	0.87(0.055)
	200		0.97(0.055)	0.98(0.005)		0.88(0.073)	0.88(0.036)
100	100	0.97	0.97(0.079)	0.97(0.009)	0.88	0.85(0.109)	0.86(0.060)
	200		0.97(0.058)	0.97(0.006)		0.86(0.076)	0.87(0.039)
500	100	0.97	0.96(0.084)	0.97(0.010)	0.87	0.83(0.117)	0.84(0.069)
	200		0.96(0.058)	0.97(0.007)		0.86(0.076)	0.86(0.038)

Table 4.2: Shrinkage factor esimates based on 1000 simulation. "Factor" indicates the theoretical asymptotic factor, "Estimate1" indicates the empirical shrinkage factor estimator, 'Estimate2" indicates the asymptotic shrinkage factor estimator. For each estimator, each entry represents mean of $1,000$ simulation results with standard error in parentheses.

		PC 1			PC 2		
γ	n	Factor	Factor Estimate1	Factor Estimate2	Factor	Factor Estimate1	Factor Estimate2
1	100	0.88	0.88(0.017)	0.87(0.076)	0.75	0.75(0.044)	0.76(0.063)
	200		0.88(0.013)	0.87(0.054)		0.75(0.027)	0.75(0.044)
20	100	0.51	0.51(0.037)	0.51(0.038)	0.33	0.34(0.033)	0.32(0.038)
	200		0.51(0.025)	0.51(0.026)		0.34(0.022)	0.33(0.028)
100	100	0.30	0.30(0.024)	0.30(0.030)	0.17	0.17(0.019)	0.17(0.023)
	200		0.30(0.017)	0.30(0.023)		0.18(0.013)	0.17(0.017)
500	100	0.16	0.15(0.014)	0.16(0.020)	0.08	0.08(0.010)	0.08(0.013)
	200		0.15(0.010)	0.16(0.014)		0.08(0.007)	0.08(0.009)

Table 4.3: Mean Square Error(MSE) of the PC regression based on gene-expression microarray data simulation with and without shrinkage adjustment. 1,000 simulation were conducted. Each entry in the table represents mean of the MSE with standard error in parentheses

n	g	Test Data without Adjustment	Test Data with Adjustment	Training Data
100	150	1.97(0.256)	1.70(0.284)	1.61(0.284)
100	300	1.63(0.230)	1.17(0.167)	1.12(0.158)
100	500	1.43(0.204)	1.07(0.157)	1.03(0.147)
100	1000	1.22(0.182)	1.03(0.148)	0.99(0.142)
200	150	1.73(0.159)	1.33(0.133)	1.30(0.131)
200	300	1.39(0.139)	1.08(0.105)	1.07(0.110)
200	500	1.24(0.131)	1.04(0.105)	1.01(0.101)
200	1000	1.10(0.114)	1.02(0.101)	1.00(0.101)

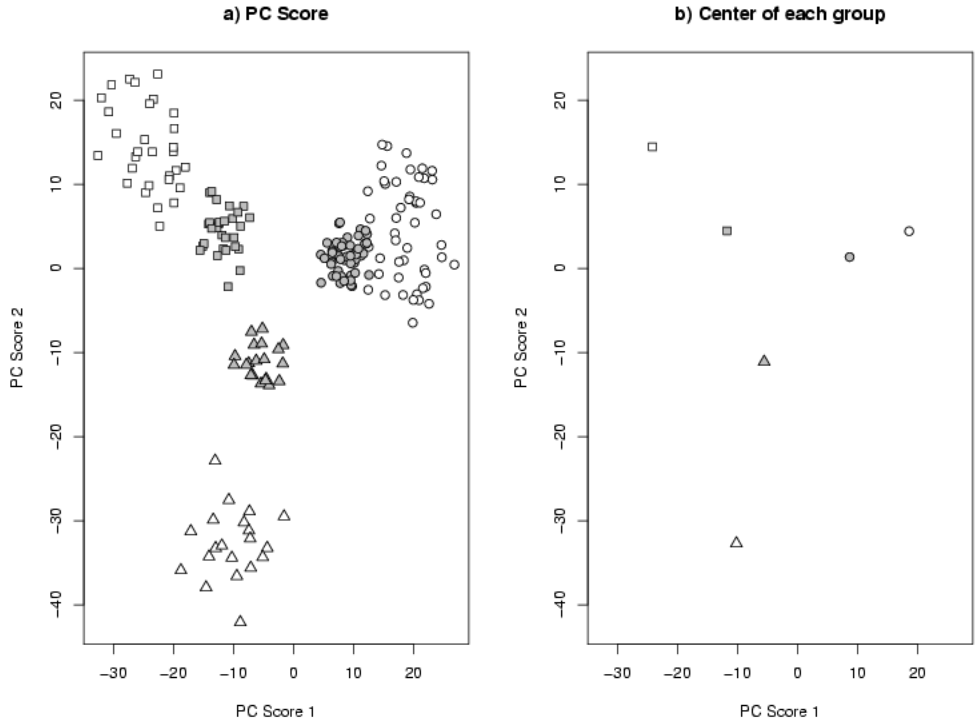

Figure 4.1: Simulation results for p=5000 and n=(50,30,20). Different symbols represent different groups. White background color represents the training set and grey background color represents the test set. A) First 2 PC score plot of all simulated samples. B) Center of each group.

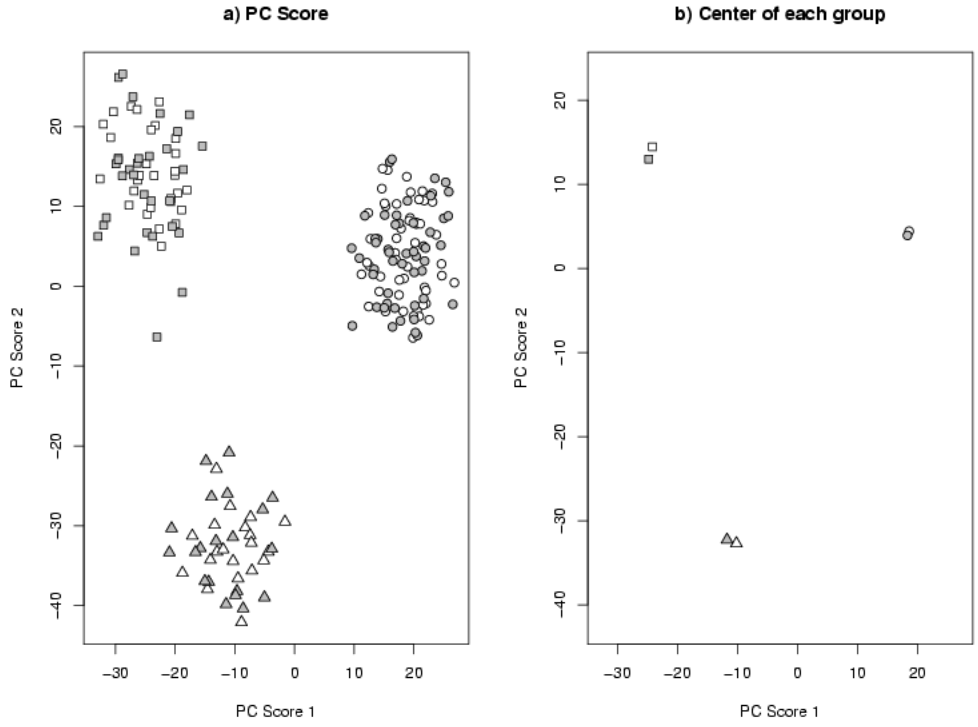

Figure 4.2: Shrinkage Adjusted PC scores of the data in Figure 1. Different symbols represent different groups. White background color represents the training set and grey background color represents the test set. A) plots of all simulation samples. B) Center of each group.

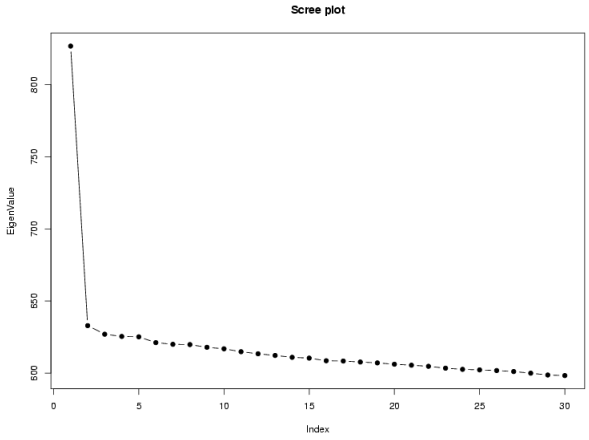

Figure 4.3: Scree plot of the first 30 sample eigenvalues, CEU+TSI dataset

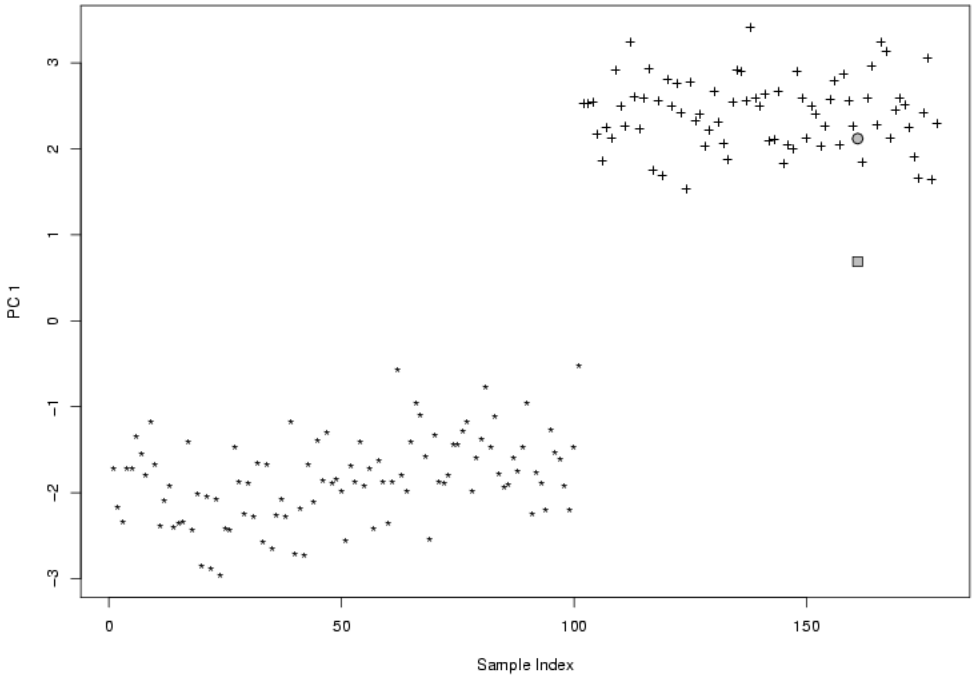

Figure 4.4: An instance with and without shrinkage adjustment, performed on Hapmap CEU(*) and TSI(+). "*" and "+" represent PC scores using all data. The 161^{th} sample was excluded from PCA, and PC score for it was predicted. The grey rectangle represents the predicted PC score without shrinkage adjustment and the grey circle represents the predicted PC score after the shrinkage adjustment

Bibliography

N.d. . Forthcoming.

Agresti, A. 2002. *Categorical Data Analysis, 2nd Edition.* Wiley New York.

Anderson, T. W. 1963. "Asymptotic theory for principal component analysis." *Ann. Math. Statist.* 34:122–148.

Anderson, T.W. 2003. *An introduction to multivariate statistical analysis, Third ed.* Wiley New York.

Bacanu, S. A., B. Devlin and K. Roeder. 2000. "The power of genomic control." *The American Journal of Human Genetics* 66(6):1933–1944.

Bacanu, S.A., B. Devlin and K. Roeder. 2002. "Association studies for quantitative traits in structured populations." *Genetic Epidemiology* 22(1):78–93.

Bai, Z. and J. Yao. 2008. Central limit theorems for eigenvalues in a spiked population model. In *Annales de lInstitut Henri Poincaré-Probabilités et Statistiques.* Vol. 44 pp. 447–474.

Bai, ZD. 1999. "Methodologies in spectral analysis of large-dimensional random matrices, a review." *Statistica Sinica* 9(3):611–677.

Baik, J., G. Ben Arous and S. Peche. 2005. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices." *Annals of probability* pp. 1643–1697.

Baik, Jinho and Jack W. Silverstein. 2006. "Eigenvalues of large sample covariance matrices of spiked population models." *J. Multivar. Anal.* 97(6):1382–1408.

Bair, E., T. Hastie, D. Paul and R. Tibshirani. 2006. "Prediction by supervised principal components." *Journal of the American Statistical Association* 101(473):119–137.

Balding, D. J. and R. A. Nichols. 1995*a*. "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity." *Genetica* 96(1):3–12.

Balding, D.J. and R.A. Nichols. 1995*b*. "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity." *Genetica* 96(1):3–12.

Bauchet, M., B. McEvoy, L. N. Pearson, E. E. Quillen, T. Sarkisian, K. Hovhannesyan, R. Deka, D. G. Bradley and M. D. Shriver. 2007. "Measuring European population stratification with microarray genotype data." *The American Journal of Human Genetics* 80(5):948–956.

Benjamini, Y. and Y. Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Bhatia, R. 1997. *Matrix analysis*. Springer Verlag.

Bovelstad, HM, S. Nygard, HL Storvold, M. Aldrin, O. Borgan, A. Frigessi and OC Lingjaerde. 2007. "Predicting survival from microarray data a comparative study." *Bioinformatics* 23:2080–2087.

Cardon, L. R. and L. J. Palmer. 2003. "Population stratification and spurious allelic association." *The Lancet* 361(9357):598–604.

Cardon, Lon R. and John I. Bell. 2001. "Association study designs for complex diseases." *Nat Rev Genet* 2(2):91–99. 10.1038/35052543.

Chatfield, C. and A. J. Collins. 1981. *Introduction to multivariate analysis*. Chapman and Hall, London.

Cheng, Q. 2002. New versions of principal component analysis for image enhancement and classification. In *Geoscience and Remote Sensing Symposium, 2002. IGARSS '02. 2002 IEEE International*. Vol. 6 pp. 3372–3374 vol.6.

Daly, A. K. and C. P. Day. 2001. "Candidate gene case-control association studies: advantages and potential pitfalls." *British Journal of Clinical Pharmacology* 52(5):489.

de Bakker, P. I. W., G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J. Monsuur, P. Whittaker and M. Delgado. 2006. "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC." *Nature genetics* 38:1166–1172.

Devlin, B. and K. Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55(4):997–1004.

Devlin, B., K. Roeder and L. Wasserman. 2001. "Genomic Control, a New Approach to Genetic-Based Association Studies." *Theoretical Population Biology* 60(3):155–166.

Devlin, B., S.A. Bacanu and K. Roeder. 2004. "Genomic Control to the extreme." *Nature Genetics* 36(11):1129–1130.

Diamantaras, K. I. and S. Y. Kung. 1996. *Principal component neural networks: theory and applications*. John Wiley & Sons, Inc. New York, NY, USA.

Easton, D. F., K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field and R. Luben. 2007. "Genome-wide association study identifies novel breast cancer susceptibility loci." *Nature* 447:1087–1095.

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

Elston, R. C. 1998. "Linkage and association." *Genetic epidemiology* 15(6):565–576.

Epstein, M.P., A.S. Allen and G.A. Satten. 2007. "A simple and improved correction for population stratification in case-control studies." *The American Journal of Human Genetics* 80(5):921–930.

Epstein, M.P., A.S. Allen and G.A. Satten. 2008. "Response to lee et al." *The American Journal of Human Genetics* 82(2):526–528.

Falush, D., M. Stephens and J. K. Pritchard. 2003. "Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies." *Genetics* 164(4):1567–1587.

Fellay, J., K.V. Shianna, D. Ge, S. Colombo, B. Ledergerber, M. Weale, K. Zhang, C. Gumbs, A. Castagna, A. Cossarizza et al. 2007. "A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1." *Science* 317(5840):944.

Fisher, R.A. 1921. "On the probable error of a coefficient of correlation deduced from a small sample." *Metron* 1(4):3–32.

Freedman, M.L., D. Reich, K.L. Penney, G.J. McDonald, A.A. Mignault, N. Patterson, S.B. Gabriel, E.J. Topol, J.W. Smoller, C.N. Pato et al. 2004. "Assessing the impact of population stratification on genetic association studies." *Nature Genetics* 36:388–393.

Girshick, MA. 1936. "Principal components." *Journal of the American Statistical Association* pp. 519–528.

Girshick, MA. 1939. "On the sampling theory of roots of determinantal equations." *The Annals of Mathematical Statistics* pp. 203–224.

Greenacre, M. J. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Hao, K., X. Xu, N. Laird and X. Wang. 2004. "Power estimation of multiple SNP association test of case-control study and application." *Genetic epidemiology* 26(1):22–30.

Hollox, E. J., M. Poulter, M. Zvarik, V. Ferak, A. Krause, T. Jenkins, N. Saha, A. I. Kozlov and D. M. Swallow. 2001. "Lactase haplotype diversity in the Old World." *The American Journal of Human Genetics* 68(1):160–172.

Horn, R.A. and C.R. Johnson. 1990. *Matrix analysis*. Cambridge Univ Pr.

Jackson, J.E. 2005. *A user's guide to principal components*. Wiley-Interscience.

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

Johnstone, I.M. and A.Y. Lu. 2007. "Sparse principal components analysis." *J. Amer. Statist. Assoc* .

Jolliffe, I.T. 2002. *Principal component analysis*. Springer New York.

Kimmel, G., M.I. Jordan, E. Halperin, R. Shamir and R.M. Karp. 2007. "A randomization test for controlling population stratification in whole-genome association studies." *The American Journal of Human Genetics* 81(5):895–905.

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

Lehmann, E.L. and J.P. Romano. 2005. *Testing Statistical Hypotheses*. Springer.

Lubin, J.H. 1981. "An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data.".

Luca, D., S. Ringquist, L. Klei, A.B. Lee, C. Gieger, H.E. Wichmann, S. Schreiber, M. Krawczak, Y. Lu, A. Styche et al. 2008. "On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants." *The American Journal of Human Genetics* 82(2):453–463.

Ma, Shuangge, Michael R. Kosorok and Jason P. Fine. 2006. "Additive risk models for survival data with high-dimensional covariates." *Biometrics* 62:202–210.

Maniatis, N., A. Collins, C. F. Xu, L. C. McCarthy, D. R. Hewett, W. Tapper, S. Ennis, X. Ke and N. E. Morton. 2002. "The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis." *Proceedings of the National Academy of Sciences* 99(4):22282233.

Marčenko, VA and LA Pastur. 1967. "DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES." *Sbornik: Mathematics* 1(4):457–483.

Marchini, J., L.R. Cardon, M.S. Phillips and P. Donnelly. 2004. "The effects of human population structure on large genetic association studies." *Nature Genetics* 36:512–517.

Miclaus, K., R. Wolfinger and W. Czika. 2009. "SNP selection and multidimensional scaling to quantify population structure." *Genetic epidemiology* .

Morrison, D.F. 1976. *Multivariate statistical methods.* McGraw-Hill, Tokyo.

Nadler, B. 2008. "Finite sample approximation results for principal component analysis: A matrix perturbation approach." *The Annals of Statistics* 36(6):2791–2817.

Patterson, N., A.L. Price and D. Reich. 2006. "Population structure and eigenanalysis." *PLoS Genet* 2(12):e190.

Paul, D. 2005. "Asymptotics of the leading sample eigenvalues for a spiked covariance model." *Technical Report* .

Paul, D. 2007. "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model." *STATISTICA SINICA* 17(4):1617.

Price, A.L., J. Butler, N. Patterson, C. Capelli, V.L. Pascali, F. Scarnicci, A. Ruiz-Linares, L. Groop, A.A. Saetta, P. Korkolopoulou et al. 2008. "Discerning the ancestry of European Americans in genetic association studies." *PLoS Genet* 4(1):e236.

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

Pritchard, J. K., M. Stephens and P. Donnelly. 2000. "Inference of population structure using multilocus genotype data." *Genetics* 155(2):945–959.

Pritchard, J. K. and N. A. Rosenberg. 1999. "Use of unlinked genetic markers to detect population stratification in association studies." *The American Journal of Human Genetics* 65(1):220–228.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker and M. J. Daly. 2007. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *The American Journal of Human Genetics* 81(3):559–575.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly et al. 2007. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *The American Journal of Human Genetics* 81(3):559–575.

Rao, N. Raj, James A. Mingo, Roland Speicher and Alan Edelman. 2008. "Statistical eigen-inference from large Wishart matrices." *Ann. Statist.* 36:2850–2885.

Reich, D.E. and D.B. Goldstein. 2001. "Detecting association in a case-control study while correcting for population stratification." *Genetic Epidemiology* 20(1):4–16.

Risch, N. and K. Merikangas. 1996. "The future of genetic studies of complex human diseases." *Science* 273(5281):1516–1517.

Sanders, A.R., J. Duan, D.F. Levinson, J. Shi, D. He, C. Hou, G.J. Burrell, J.P. Rice, D.A. Nertney, A. Olincy et al. 2008. "No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics." *American Journal of Psychiatry* 165:497–506.

Satten, G. A., W. D. Flanders and Q. Yang. 2001. "Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model." *The American Journal of Human Genetics* 68(2):466–477.

Saxena, R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn and M. J. Daly. 2007. "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." *Science* 316(5829):1331–1336.

Schork, N. J., D. Fallin, B. Thiel, X. Xu, U. Broeckel, H. J. Jacob and D. Cohen. 2001. "The future of genetic case-control studies." *Adv Genet* 42:191–212.

Schulze, T. G., F. J. McMahon and F. B. Methods. 2002. "Invited Paper Genetic Association Mapping at the Crossroads: Which Test and Why? Overview and Practical Guidelines." *American Journal of Medical Genetics (Neuropsychiatric Genetics)* 114:1–11.

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines and A. U. Jackson. 2007. "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants." *Science* 316(5829):1341–1345.

Service, Susan, Joseph DeYoung, Maria Karayiorgou, J. Louw Roos, Herman Pretorious, Gabriel Bedoya, Jorge Ospina, Andres Ruiz-Linares, Antonio Macedo, Joana Almeida Palha, Peter Heutink, Yurii Aulchenko, Ben Oostra, Cornelia van Duijn, Marjo-Riitta Jarvelin, Teppo Varilo, Lynette Peddle, Proton Rahman, Giovanna Piras, Maria Monne, Sarah Murray, Luana Galver, Leena Peltonen, Chiara Sabatti, Andrew Collins and Nelson Freimer. 2006. "Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies." *Nat Genet* 38(5):556–560. 10.1038/ng1770.

Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle and S. Hadjadj. 2007. "A genome-wide association study identifies novel risk loci for type 2 diabetes." *Nature* 445:881–885.

Stefansson, H., A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason and V. G. Gudnadottir. 2005. "A common inversion under selection in Europeans." *Nature genetics* 37:129–137.

Stirling, W.D. 1982. "Enhancements to aid interpretation of probability plots." *The Statistician* 31:211–20.

Tian, C., R. Kosoy, A. Lee, M. Ransom, J.W. Belmont, P.K. Gregersen and M.F. Seldin. 2008. "Analysis of East Asia genetic substructure using genome-wide SNP arrays." *Plos One* 3(12).

Wall, M.E., A. Rechtsteiner and L.M. Rocha. 2003. "Singular value decomposition and principal component analysis." *A practical approach to microarray data analysis* pp. 91–109.

Wright, F. A., H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, P. M. de Villena, P. F. Sullivan, K. C. Wilhelmsen and F. Zou. 2007. "Simulating association studies: a data-based resampling method for candidate regions or whole genome scans." *Bioinformatics* 23(19):2581–2588.

Zeggini, E., M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. B. Perry, N. W. Rayner and R. M. Freathy. 2007. "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes." *Science* 316(5829):1336–1341.

Zhang, Feng, Yuping Wang and Hong-Wen Deng. 2008. "Comparison of Population-Based Association Study Methods Correcting for Population Stratification." *PLoS ONE* 3(10):e3392.

Zhang, S., X. Zhu and H. Zhao. 2003. "On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals." *Genetic epidemiology* 24(1):44–56.

Zhao, H., T.R. Rebbeck and N. Mitra. 2009. "A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors." *to appear in Genetic Epidemiology* .

Zhu, X., S. L. Zhang, H. Zhao and R. S. Cooper. 2002. "Association mapping, using a mixture model for complex traits." *Genetic epidemiology* 23(2):181–196.