The Annals of Statistics
2010, Vol. 38, No. 6, 3605–3629
DOI: 10.1214/10-AOS821

CONVERGENCE AND PREDICTION OF PRINCIPAL COMPONENT SCORES IN HIGH-DIMENSIONAL SETTINGS

BY SEUNGGEUN LEE[1], FEI ZOU[1,2] AND FRED A. WRIGHT[1,2]

University of North Carolina at Chapel Hill

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 toward 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.

1. Introduction. Principal component analysis (PCA) [19] 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 [24, 27]. 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 [30].

Although PCA is widely applied in a number of settings, much of our theoretical understanding rests on a relatively small body of literature. Girshick [12] 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

Received December 2009; revised February 2010.
[1]Supported in part by NIH Grant GM074175.
[2]Supported in part by the Carolina Environmental Bioinformatics Center (EPA RD832720) and a Gillings Innovation Award.
AMS 2000 subject classifications. Primary 62H25; secondary 15A18.
Key words and phrases. PCA, PC scores, random matrix, PC regression.

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 [2, 13]. 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 [5]). The convergence of the sample eigenvalues and eigenvectors under the "spiked population" model proposed by Johnstone [18] has also been established [7, 23, 26]. 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) [17, 23, 26]. Furthermore, Paul [26] 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 [23] 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 subtype. The recalculation of PCs using both new and old data might not be practical. For example, 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 [16]. 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 toward 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.

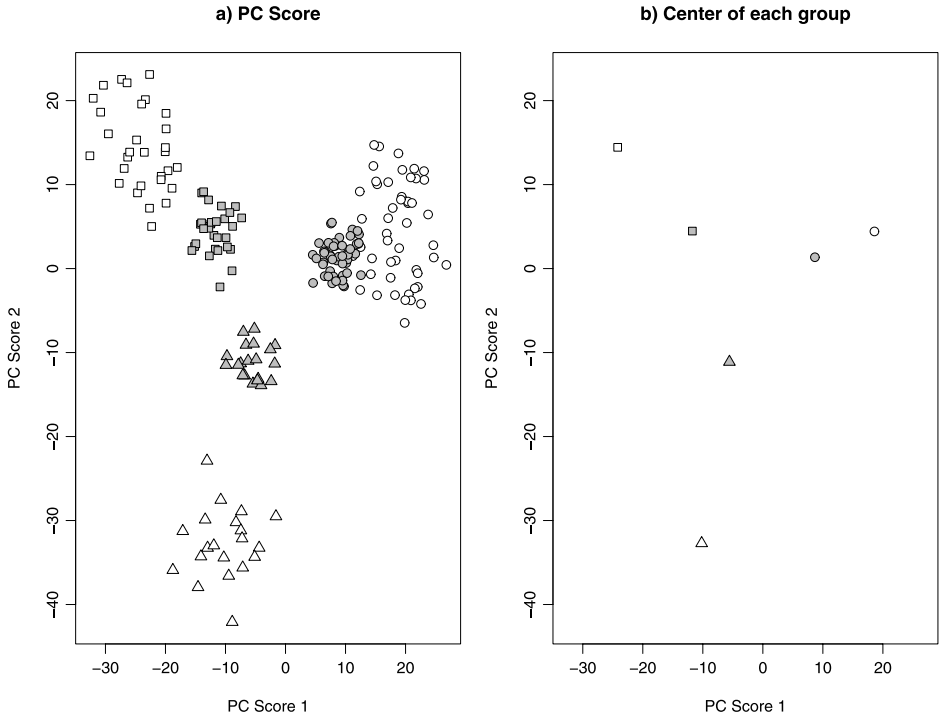

FIG. 1. *Simulation results for* $p = 5000$ *and* $n = (50, 30, 20)$. *Different symbols represent different groups. White background color represents the training set and grey background color represents the test set.* (a) *First* 2 *PC score plot of all simulated samples.* (b) *Center of each group.*

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 $\text{MVN}(\boldsymbol{\mu}_k, 4\mathbf{I})$, 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 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 and grow at the same rate. 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 [26], 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 2.

In Section 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, 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.

2. Method.

2.1. *General setting*. Throughout this paper, we use ^{T} to denote matrix transpose, $\xrightarrow{p}$ to denote convergence in probability, and $\xrightarrow{\text{a.s.}}$ to denote almost sure convergence. Let $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, 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 jth 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_k's are the underlying population eigenvalues. The spiked population model defined in [18] assumes that all the population eigenvalues are 1, except the first m eigenvalues. That is, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > \lambda_{m+1} = \cdots = \lambda_p = 1$. The spectral decomposition of the sample covariance matrix is

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

where $\mathbf{D} = \operatorname{diag}(d_1, d_2, \ldots, d_p)$ 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 $\mathbf{x}_{\text{new}}$, its predicted PC score is similarly defined as $\mathbf{U}^T \mathbf{x}_{\text{new}}$ with the vth (PC) score equal to $q_v = \mathbf{u}_v^T \mathbf{x}_{\text{new}}$.

2.2. *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 [3]. 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 $\gamma = 0$,*

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

(ii) *When $\gamma > 0$,*

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

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 [7], while the proof of (i) can be found in Section 6.3. 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 nonzero γ, d_v is no longer a consistent estimator of λ_v. However, a consistent estimator of λ_v can be constructed from (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 Péché [6] have shown the $\sqrt{n}$-consistency of d_v to $\rho(\lambda_v)$, and Bai and Yao [4] have shown that d_v is asymptotically normal.

LEMMA 2. *Suppose $p/n \to \gamma \geq 0$ as $n \to \infty$. Let $\langle \cdot, \cdot \rangle$ 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*

$$|\langle \mathbf{e}_v, \mathbf{u}_v \rangle| \xrightarrow{a.s.} \begin{cases} \phi(\lambda_v), & \text{if } \lambda_v > 1 + \sqrt{\gamma}, \\ 0, & \text{if } 1 < \lambda_v \leq 1 + \sqrt{\gamma}; \end{cases} \tag{3}$$

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

$$|\langle \mathbf{e}_v, \mathbf{u}_v \rangle| \xrightarrow{p} \begin{cases} \phi(\lambda_v), & \text{if } \lambda_v > 1 + \sqrt{\gamma}, \\ 0, & \text{if } 1 < \lambda_v \leq 1 + \sqrt{\gamma}. \end{cases} \tag{4}$$

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 [26] proved it for $\gamma < 1$; while Nadler [23] 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 6.4. 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.

2.3. *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 the following theorem.

THEOREM 1. *Let $\mathbf{g}_v^T = \mathbf{e}_v^T \mathbf{X}/\sqrt{\mathbf{n}\lambda_v}$, the normalized vth PC score derived from a corresponding population eigenvector, $\mathbf{e}_v$, and $\tilde{\mathbf{p}}_v = \mathbf{p}_v/\sqrt{\mathbf{n}d_v}$, the normalized vth sample PC score. Suppose $p/n \to \gamma \geq 0$ as $n \to \infty$. Under the multiplicity one assumption,*

$$|\langle \mathbf{g}_v, \tilde{\mathbf{p}}_v \rangle| \xrightarrow{p} \begin{cases} \sqrt{1 - \dfrac{\gamma}{(\lambda_v - 1)^2}}, & \text{if } \lambda_v > 1 + \sqrt{\gamma}, \\ 0, & \text{if } 1 < \lambda_v \leq 1 + \sqrt{\gamma}. \end{cases} \tag{5}$$

The proof can be found in Section 6.7. 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 $\langle \mathbf{g}_v, \tilde{\mathbf{p}}_v \rangle$ is the correlation coefficient between $\mathbf{g}_v$ and $\tilde{\mathbf{p}}_v$. Compared to (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

$$\mathbf{e}_1 = \underset{\mathbf{a}:\, \mathbf{a}^T\mathbf{a}=1}{\arg\max}\, E((\mathbf{a}^T\mathbf{x})^2)$$

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

$$\mathbf{u}_1 = \arg\max_{\mathbf{a}:\,\mathbf{a}^T\mathbf{a}=1} \sum_{j=1}^{n} (\mathbf{a}^T \mathbf{x}_j)^2.$$

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

$$\begin{aligned} E((\mathbf{u}_1^T \mathbf{x}_{\text{new}})^2) &= E(E(\mathbf{u}_1^T \mathbf{x}_{\text{new}} \mathbf{x}_{\text{new}}^T \mathbf{u}_1^T | \mathbf{u}_1)) = E(\mathbf{u}_1^T E(\mathbf{x}_{\text{new}} \mathbf{x}_{\text{new}}^T) \mathbf{u}_1^T) \\ &= E(\mathbf{u}_1^T \boldsymbol{\Sigma} \mathbf{u}_1^T) \le \mathbf{e}_1^T \boldsymbol{\Sigma} \mathbf{e}_1 = E((\mathbf{e}_1^T \mathbf{x}_{\text{new}})^2). \end{aligned} \tag{6}$$

Since the $\mathbf{u}_1^T \mathbf{x}_j$'s ($j = 1, \ldots, n$) follow the same distribution,

$$nE((\mathbf{e}_1^T \mathbf{x}_j)^2) = E\left(\sum_{j=1}^{n} (\mathbf{e}_1^T \mathbf{x}_j)^2\right) \le E\left(\sum_{j=1}^{n} (\mathbf{u}_1^T \mathbf{x}_j)^2\right) = nE((\mathbf{u}_1^T \mathbf{x}_j)^2). \tag{7}$$

By (6) and (7), we can show that

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

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

$$\sqrt{\frac{E(q_v^2)}{E(p_{vj}^2)}} \xrightarrow{n\to\infty} \frac{\lambda_v - 1}{\lambda_v + \gamma - 1}, \tag{8}$$

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

The proof is given in Section 6.8. 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.

2.4. *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 $\boldsymbol{\Lambda}^* = \{\zeta\lambda_1, \ldots, \zeta\lambda_m, \zeta, \ldots, \zeta\}$, where $\zeta \ne 1$, and its corresponding sample eigenvalues $\mathbf{D}^* = \{d_1^*, \ldots,$

$d_n^*\}$. We can show that (4), (8) and (5) still hold under such circumstances. However, $\rho^{-1}(d_v^*)$ is no longer a consistent estimator of λ_v, because

$$d_v^* \xrightarrow{\text{a.s.}} \zeta\lambda_v\left(1+\frac{\gamma}{\lambda_v-1}\right)=\zeta\rho(\lambda_v).$$

To address this issue, Baik and Silverstein [7] 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 Marčenko–Pastur law [22]. 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

$$E\left(\sum_{v=1}^{p} d_v\right)=E(\text{trace}(\mathbf{S}))=\text{trace}(E(\mathbf{S}))=\text{trace}(\mathbf{\Sigma})=\sum_{v=1}^{p}\lambda_v.$$

Thus, the sum of the rescaled eigenvalues is expected to be close to $(\sum_{v=1}^{m}\lambda_v + p-m)$. Let $r_v = d_v^*/(\sum_{v=1}^{p} d_v^*)$ 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_v's 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 lth 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 $\hat{d}_{v,l-1} \le (1+\sqrt{\gamma})^2$. 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 3. *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 $\tilde{\mathbf{p}}_v$

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

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 2. After the bias adjustment, the predicted PC scores of the test data are comparable to those of the training data.

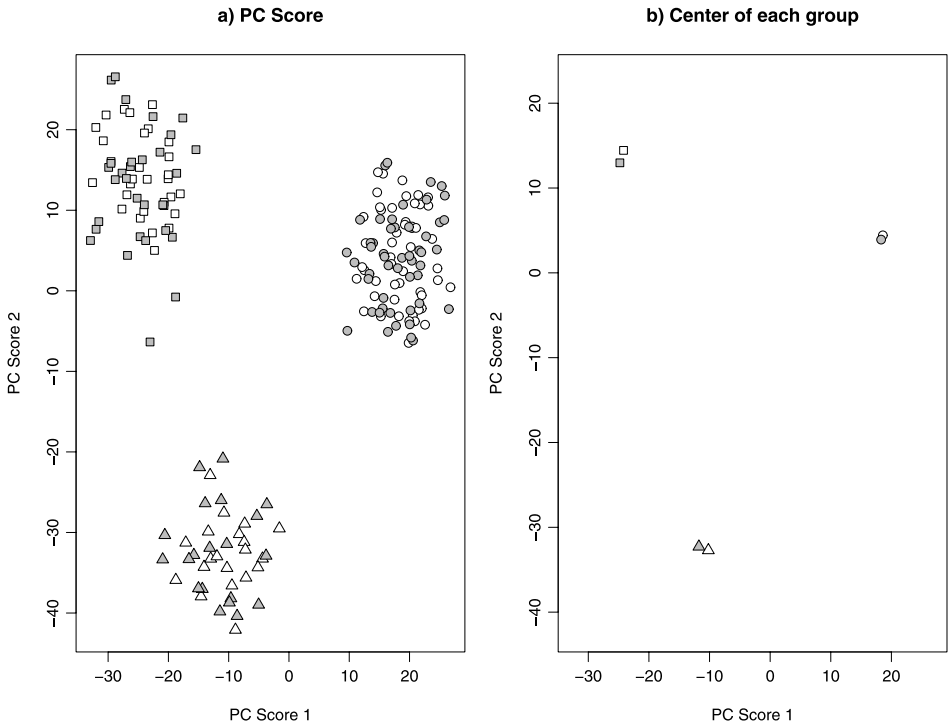

FIG. 2. *Shrinkage-adjusted PC scores of the data in Figure 1. Different symbols represent different groups. White background color represents the training set and grey background color represents the test set.* (a) *Plots of all simulation samples.* (b) *Center of each group.*

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 ith 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 1 and 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 [9]. In this simulation, we let $p = 5000$, and our set up is very similar to the first simulation of Bair et al. [8]. Let x_{ij} denote the gene expression level of the ith gene for the jth subject. We generated each x_{ij} according to

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

and the outcome variable y_j as

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

where n is the number of samples, g is the number of genes that are differentially expressed and associated with the phenotype, $\varepsilon \sim N(0, 2^2)$ and $\varepsilon_y \sim N(0, 1)$. A total of eight different combinations of n and g were simulated. For the training data,

TABLE 1

Cosine angle estimates of eigenvectors and PC scores based on 1000 *simulations. "Angle" indicates the theoretical asymptotic* cos(angle), *"Estimate*1*" indicates the empirical* cos(angle) *estimator, "Estimate*2*" indicates the asymptotic* cos(angle) *estimator. For each estimator, each entry represents mean of* 1000 *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)

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 unadjusted and adjusted PC scores. The results are presented in Table 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. 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 an-

TABLE 2

Shrinkage factor estimates based on 1000 *simulation. "Factor" indicates the theoretical asymptotic factor, "Estimate*1*" indicates the empirical shrinkage factor estimator, "Estimate*2*" indicates the asymptotic shrinkage factor estimator. For each estimator, each entry represents mean of* 1000 *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)

cestry) 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) [11]. Lastly, we excluded 7 samples with PC scores greater than 6 standard deviations away from the mean of at least one of the top

TABLE 3

Mean Square Error (MSE) of the PC regression based on gene-expression microarray data simulation with and without shrinkage adjustment. 1000 *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)

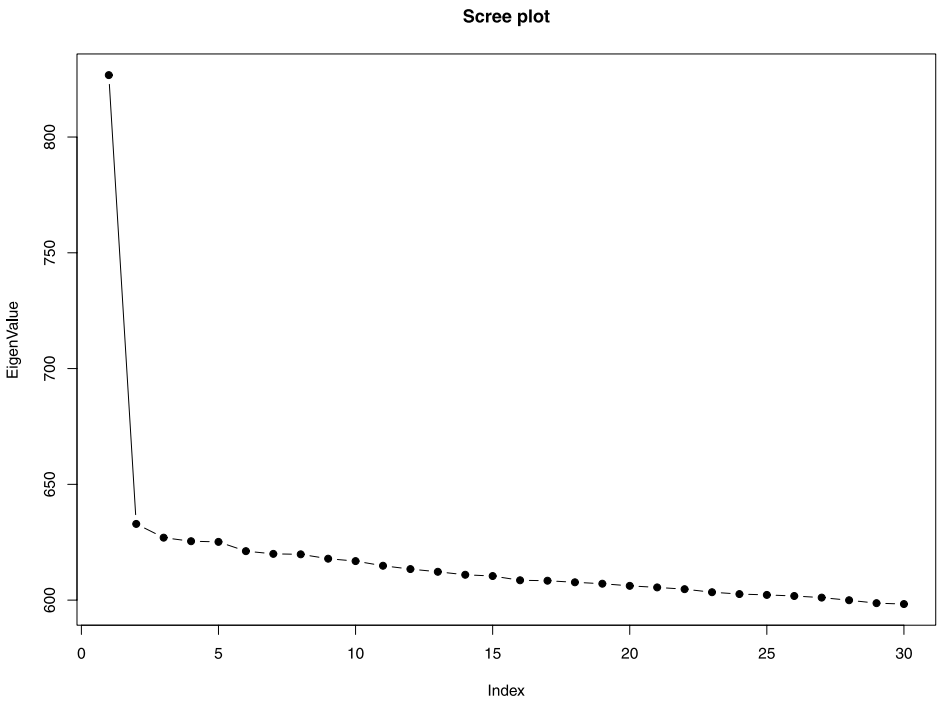

FIG. 3. *Scree plot of the first* 30 *sample eigenvalues*, *CEU* + *TSI dataset*.

significant PCs [i.e., with Tracy–Widom (TW) Test p-value <0.01] [24, 27]. 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 [27]. The screeplot of the sample eigenvalues is presented in Figure 3. The first eigenvalue is substantially larger than the rest of the eigenvalues, although the TW test actually identifies two significant PCs. Figure 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 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 toward zero, while the bias-adjusted PC score is not.

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

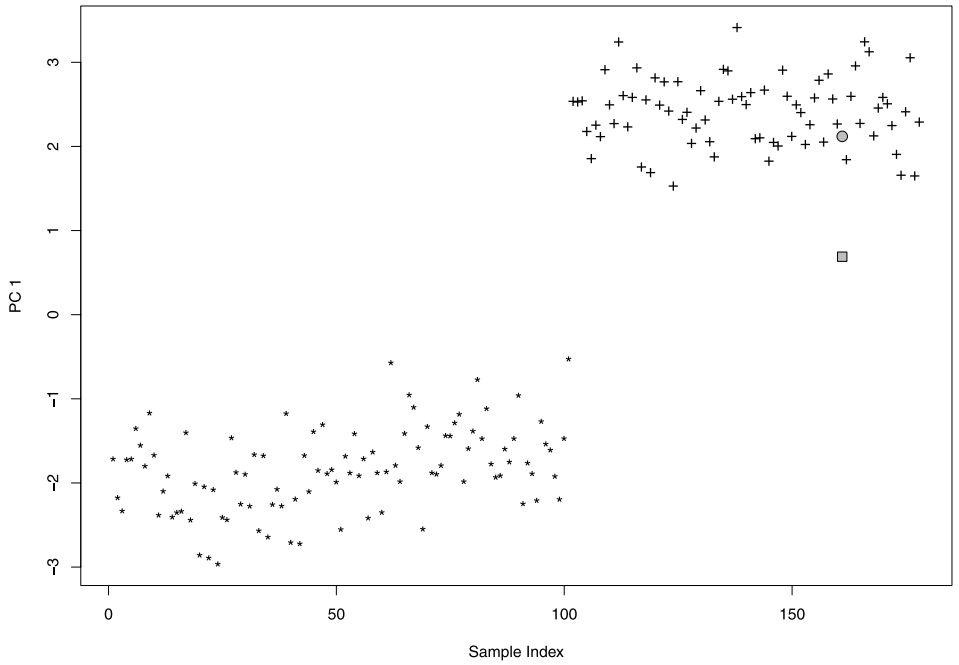

FIG. 4. *An instance with and without shrinkage adjustment, performed on Hapmap CEU(*) and TSI(+). "*" and "+" represent PC scores using all data. The* 161*st sample was excluded from PCA, and PC score for it was predicted. The grey rectangle represents the predicted PC score without shrinkage adjustment and the grey circle represents the predicted PC score after the shrinkage adjustment.*

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.

We believe that this asymptotic regime applies well to many high-dimensional datasets. It is not, however, the only model paradigm applied to such data. For example, the large p small n paradigm [1, 14], which assumes $p/n \to \infty$, has also been explored. Under this assumption, Jung and Marron [20] have shown that the consistency and the strong inconsistency of the sample eigenvectors to population eigenvectors depend on whether p increases at a slower or faster rate than λ_v. It may be argued that for real data where p/n is "large," we should follow the paradigm of Hall, Marron and Neeman [14], Ahn et al. [1]. However, for any real study, it is unclear how to test whether p increases at a faster rate than λ_v, or vice versa, making the application of Hall, Marron and Neeman [14], Ahn et al. [1] dif-

ficult in practice. Furthermore, the scenario where p and λ_v grow at the same rate is scientifically more interesting, for which we are aware of no theoretical results. In contrast, our asymptotic results can be straightforwardly applied. Further, our simulation results indicate that for p/n as large as 500, our asymptotic results still hold well. We believe that the approach we describe here applies to many datasets.

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 [10, 29]. 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 [27], and in PC regression for microarrays [21]. 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 [28]. 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.

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.

6.1. *Notation*. We largely follow notation in Paul [26]. We denote $\lambda_v(\mathbf{S})$ as the vth 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 vth 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

$$\frac{1}{\sqrt{n}}\mathbf{Z}_B = \mathbf{V}\mathbf{M}^{1/2}\mathbf{H}^T, \tag{9}$$

where $\mathbf{M} = \operatorname{diag}(\mu_1, \ldots, \mu_{p-m})$ 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.

6.2. *Propositions.* We introduce two propositions for later use. The proofs of the two propositions can be found in Sections 6.5 and 6.6.

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 nonnegative definite random matrix and independent of* $\mathbf{Y}$. *Further, assume* $\|\mathbf{C}\| = O(1)$. *Then*

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

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* $\mathbf{F} = \operatorname{diag}(f(\mu_1), \ldots, f(\mu_{p-m}))$, *where* $\{\mu_i\}_{i=1}^{p-m}$ *are ordered eigenvalues of* $\mathbf{M}$ *which is defined on* (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 Marčenko–Pastur law with parameter* γ [22].

6.3. *Proof of part* (i) *of Lemma* 1.

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

6.3.2. *When* $p \to \infty$. For every small $\varepsilon > 0$, there exist $\tilde{p}(n)$ and γ_ε such that $\tilde{p}(n)/n \to \gamma_\varepsilon > 0$, $\lambda_v(1 + \gamma_\varepsilon/(\lambda_v - 1)) < \lambda_v + \varepsilon$ for all $v \le m$, $(1 + \sqrt{\gamma_\varepsilon})^2 < 1 + \varepsilon$ and $(1 - \sqrt{\gamma_\varepsilon})^2 > 1 - \varepsilon$. 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 $v \le (m+1)$, by the interlacing inequality (Theorem 4.3.15 of Horn and Johnson [15]),

$$\lambda_v(\hat{\mathbf{S}}) \le \lambda_v(\mathbf{S}) \le \lambda_v(\tilde{\mathbf{S}}).$$

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

$$\lambda_v - o(1) \le \lambda_v(\mathbf{S}) < \lambda_v + \varepsilon + o(1) \qquad \text{for } v \le m + 1.$$

Thus,

$$\lambda_v(\mathbf{S}) \xrightarrow{\text{a.s.}} \lambda_v \qquad \text{for } v \le m + 1. \tag{10}$$

Similarly by the interlacing inequality, we get

$$\lambda_{\tilde{p}}(\tilde{S}) \le \lambda_p(S) \le \lambda_{m+1}(S).$$

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

$$\lambda_p(S) \xrightarrow{\text{a.s.}} 1. \tag{11}$$

Part (i) of Lemma 1 follows by (10) and (11).

6.4. *Proof of part* (ii) *of Lemma 2*. Our proof of Lemma 2(ii) closely follows the arguments in Paul [26]. From [26], it can be shown that

$$\left(\mathbf{S}_{AA} + \frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(d_v\mathbf{I} - \mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2}\right)\mathbf{a}_v = d_v\mathbf{a}_v \tag{12}$$

and

$$\mathbf{a}_v^T\left(\mathbf{I} + \frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(d_v\mathbf{I} - \mathbf{M})^{-2}\mathbf{H}^T\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2}\right)\mathbf{a}_v = \frac{1}{1 - R_v^2}, \tag{13}$$

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

6.4.1. *When* $\lambda_v > 1 + \sqrt{\gamma}$. We can show that

$$\langle \mathbf{a}_v, \mathbf{e}_{A,v} \rangle \xrightarrow{p} 1 \tag{14}$$

and

$$\frac{1}{n}\mathbf{z}_{Av}^T\mathbf{H}\mathbf{M}(d_v\mathbf{I} - \mathbf{M})^{-2}\mathbf{H}^T\mathbf{z}_{Av} \xrightarrow{p} \begin{cases} \gamma \displaystyle\int \frac{x}{(\rho_v - x)^2}\, dF_\gamma(x), & \text{for } \gamma > 0, \\ 0, & \text{for } \gamma = 0, \end{cases} \tag{15}$$

where $\mathbf{e}_{A,v}$ is a vector of the first m coordinates of the vth population eigenvector $\mathbf{e}_v$, ρ_v is $\lambda_v(1+\frac{\gamma}{\lambda_v-1})$ and $\mathbf{z}_{Av}$ is a vector of vth row of $\mathbf{Z}_A$. The proofs can be found in Section 6.4.3. 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 (13) and (15), we can show that

$$\frac{1}{1-R_v^2} \xrightarrow{p} \begin{cases} 1+\lambda_v\gamma \displaystyle\int \frac{x}{(\rho_v-x)^2}\,dF_\gamma(x), & \text{for } \gamma>0, \\ 1, & \text{for } \gamma=0. \end{cases} \tag{16}$$

From Lemma B.2 of [26],

$$\int \frac{x}{(\rho_v-x)^2}\,dF_\gamma(x) = \frac{1}{(\lambda_v-1)^2-\gamma}. \tag{17}$$

Thus,

$$\sqrt{1-R_v^2} \xrightarrow{p} \begin{cases} \sqrt{\left(1-\dfrac{\gamma}{(\lambda_v-1)^2}\right)\Big/\left(1+\dfrac{\gamma}{\lambda_v-1}\right)}, & \text{for } \gamma>0, \\ 1, & \text{for } \gamma=0. \end{cases} \tag{18}$$

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

6.4.2. *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 $\varepsilon>0$ and $x\geq 0$, define

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

and

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

then by Propositions 1 and 2,

$$\frac{1}{n}\mathbf{z}_{Av}^T\mathbf{H}\mathbf{G}_\varepsilon\mathbf{H}^T\mathbf{z}_{Av} \xrightarrow{p} \gamma\int \frac{x}{((\rho_v-x)^2)_\varepsilon}\,dF_\gamma(x). \tag{19}$$

By monotone convergence theorem,

$$\gamma\int \frac{x}{((\rho_v-x)^2)_\varepsilon}\,dF_\gamma(x) \xrightarrow{\varepsilon\to 0} \gamma\int \frac{x}{(\rho_v-x)^2}\,dF_\gamma(x). \tag{20}$$

The right-hand side of (20) is

$$\int_a^b \frac{\sqrt{(b-x)(x-a)}}{2\pi(\rho_v-x)^2}\,dx, \tag{21}$$

where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$. Since (21) equals ∞ for any $a \le \rho_v \le b$, we conclude that

$$\frac{1}{n}\mathbf{z}_{Av}^T\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-2}\mathbf{H}^T\mathbf{z}_{Av} \xrightarrow{p} \infty. \tag{22}$$

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

6.4.3. *Proof of* (14) *and* (15). Define

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

$$\mathcal{D}_v = \mathbf{S}_{AA} + \mathbf{S}_{AB}(d_v\mathbf{I}-\mathbf{S}_{BB})^{-1}\mathbf{S}_{BA} - (\rho_v/\lambda_v)\mathbf{\Lambda}_A,$$

$$\alpha_v = \|\mathcal{R}_v\mathcal{D}_v\| + |d_v-\rho_v|\|\mathcal{R}_v\| \quad \text{and} \quad \beta_v = \|\mathcal{R}_v\mathcal{D}_v\mathbf{e}_{A,v}\|.$$

With the exactly same argument of [26], 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 [25], $r_v = o_p(1)$, 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

$$\mathbf{S}_{AB}(d_v\mathbf{I}-\mathbf{S}_{BB})^{-1}\mathbf{S}_{BA} = \frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(d_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A^T\mathbf{\Lambda}_A^{1/2}. \tag{23}$$

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

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

By Proposition 1,

$$0 \le \|(23)\| \le \lambda_1\frac{p\mu_1}{n(d_v-\mu_1)} + o_p(1) = o_p(1), \tag{24}$$

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

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

$$\begin{aligned}\mathcal{D}_v = {}& [\mathbf{S}_{AA}-\mathbf{\Lambda}_A] \\ &+ \Big[\mathbf{\Lambda}_A^{1/2}\Big(\frac{1}{n}\mathbf{Z}_A\mathbf{H}\mathbf{M}(\rho_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A \\ &\qquad - \frac{1}{n}\operatorname{trace}\big(\mathbf{M}(\rho_v\mathbf{I}-\mathbf{M})^{-1}\big)\mathbf{I}\Big)\mathbf{\Lambda}_A^{1/2}\Big] \\ &+ \Big[\Big(\frac{1}{n}\operatorname{trace}\big(\mathbf{M}(\rho_v\mathbf{I}-\mathbf{M})^{-1}\big) - \gamma\int\frac{x}{\rho_v-x}\,dF_\gamma(x)\Big)\mathbf{\Lambda}_A\Big] \\ &+ \Big[(\rho_v-d_v)\frac{1}{n}\mathbf{\Lambda}_A^{1/2}\mathbf{Z}_A\mathbf{H}\mathbf{M}(\rho_v\mathbf{I}-\mathbf{M})^{-1}(d_v\mathbf{I}-\mathbf{M})^{-1}\mathbf{H}^T\mathbf{Z}_A\mathbf{\Lambda}_A^{1/2}\Big].\end{aligned} \tag{25}$$

The first term of the right-hand side 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 (14).

For (15): When $\gamma = 0$, (15) can be proved by the exactly same way used to show (24). When $\gamma > 0$, $d_v \xrightarrow{\text{a.s.}} \rho_v$, and $\mu_1 \xrightarrow{\text{a.s.}} (1+\sqrt{\gamma})^2 < \rho_v$, hence $\|\mathbf{C}\| \xrightarrow{\text{a.s.}} \frac{(1+\sqrt{\gamma})^2}{(\rho_v-(1+\sqrt{\gamma})^2)^2}$. Therefore, the result follows according to Propositions 1 and 2.

6.5. *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 $\psi(s,s) = \frac{1}{n}\mathbf{y}_s^T\mathbf{C}\mathbf{y}_s - \frac{1}{n}\operatorname{trace}(\mathbf{C})$ 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

$$
\begin{aligned}
E(\psi(s,s)|\mathbf{C}) &= E\Big(\frac{1}{n}\sum_{i,j} c_{ij}y_{is}y_{js}|\mathbf{C}\Big) - \frac{1}{n}\sum_{i=1}^n \mu_i \\
&= \frac{1}{n}\sum_{i=1}^n c_{ii}E(y_{is}^2) + \frac{2}{n}\sum_{i<j}^n c_{ij}E(y_{is}y_{js}) - \frac{1}{n}\sum_{i=1}^n \mu_i \\
&= \frac{1}{n}\sum_{i=1}^n c_{ii} - \frac{1}{n}\sum_{i=1}^n \mu_i = 0.
\end{aligned} \tag{26}
$$

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

$$
\begin{aligned}
\operatorname{Var}(\psi(s,s)|\mathbf{C}) &= \frac{1}{n^2}\operatorname{Var}\Big(\sum_{i,j} c_{ij}y_{is}y_{js}|\mathbf{C}\Big) \\
&= \frac{1}{n^2}\sum_{i,j,l,q=1}^n c_{ij}c_{lq}\operatorname{Cov}(y_{is}y_{js}, y_{ls}y_{qs}) \\
&= \frac{4}{n^2}\sum_{i,j=1}^n c_{ij}^2\operatorname{Var}(y_{is}y_{js}) \leq \frac{4\alpha}{n^2}\sum_{i,j=1}^n c_{ij}^2 \\
&= \frac{4\alpha}{n^2}\operatorname{trace}(\mathbf{C}^2) = \frac{4\alpha}{n^2}\sum_{i=1}^n \mu_i^2,
\end{aligned} \tag{27}
$$

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

6.6. *Proof of Proposition 2.* Consider an expansion

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

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

(a): Since $\mu_1 \xrightarrow{\text{a.s.}} (1+\sqrt{\gamma})^2$, $\mu_{\min(p-m,n)} \xrightarrow{\text{a.s.}} (1-\sqrt{\gamma})^2$, $\mu_k = 0$ 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 $\mathbf{C} = \mathbf{H}\mathbf{F}\mathbf{H}^T$, then $\operatorname{trace}(\mathbf{C}) = \operatorname{trace}(\mathbf{F})$. By Proposition 1, (a) $= o_p(1)$.

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

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

and $\int f(x)\, dF_n(x) \xrightarrow{p} \int f(x)\, dF_\gamma(x)$ [5, 22]. 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.

6.7. *Proof of Theorem 1.* Without loss of generality, 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 vth 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

$$\begin{aligned}
\langle \mathbf{g}_v, \tilde{\mathbf{p}}_v \rangle &= \frac{1}{n\sqrt{d_v \lambda_v}} \mathbf{e}_v^T \mathbf{X}\mathbf{X}^T \mathbf{u}_v \\
&= \mathbf{e}_{A,v}^T \mathbf{S}_{AA}\mathbf{u}_{A,v}/\sqrt{d_v\lambda_v} + \mathbf{e}_{A,v}^T \mathbf{S}_{AB}\mathbf{u}_{B,v}/\sqrt{d_v\lambda_v} \\
&= \frac{d_v}{\sqrt{d_v\lambda_v}} \mathbf{e}_{A,v}^T \mathbf{u}_{A,v} = \sqrt{\frac{d_v}{\lambda_v}}\mathbf{e}_v^T \mathbf{u}_v \\
&\xrightarrow{p} \begin{cases} \sqrt{\left(1 - \dfrac{\gamma}{(\lambda_v - 1)^2}\right)}, & \text{for } \lambda_v > 1 + \sqrt{\gamma}, \\ 0, & \text{for } 1 < \lambda_v \leq 1 + \sqrt{\gamma}. \end{cases}
\end{aligned} \tag{28}$$

6.8. *Proof of Theorem 2.* 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 $i \neq j$,

$$\begin{aligned} E(p_{vj}^2) &= \frac{1}{n} E\left(\sum_{j=1}^{n} p_{vj}^2\right) = \frac{1}{n} E\left(\sum_{j=1}^{n} (\mathbf{u}_v^T \mathbf{x}_j)^2\right) \\ &= E(\mathbf{u}_v^T \mathbf{X}\mathbf{X}^T \mathbf{u}_v / n) = E(d_v) \xrightarrow{\text{a.s.}} \rho(\lambda_v). \end{aligned} \tag{29}$$

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 $\mathbf{u}_v^{\perp} = \{\mathbf{u}_{A,v}^{\perp}, \mathbf{u}_{B,v}^{\perp}\}$. From (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 $\mathbf{x}_{\text{new}}$ and $\mathbf{u}_v$ are independent, we have

$$\begin{aligned} E(q_v^2|\mathbf{u}_v) &= E((\mathbf{u}_v^T \mathbf{x}_{\text{new}})^2|\mathbf{u}_v) = \mathbf{u}_v^T E(\mathbf{x}_{\text{new}}\mathbf{x}_{\text{new}}^T|\mathbf{u}_v)\mathbf{u}_v = \mathbf{u}_v^T \mathbf{\Lambda}\mathbf{u}_v \\ &= (\mathbf{u}_v^T \mathbf{e}_v)^2 \mathbf{e}_v^T \mathbf{\Lambda}\mathbf{e}_v + \big(1 - (\mathbf{u}_v^T \mathbf{e}_v)^2\big)\mathbf{u}_v^{\perp T}\mathbf{\Lambda}\mathbf{u}_v^{\perp} \\ &\quad + 2\mathbf{u}_v^T \mathbf{e}_v \sqrt{1 - (\mathbf{u}_v^T \mathbf{e}_v)^2}\mathbf{e}_v^T \mathbf{\Lambda}\mathbf{u}_v^{\perp} \\ &= (\mathbf{u}_v^T \mathbf{e}_v)^2 \lambda_v + \big(1 - (\mathbf{u}_v^T \mathbf{e}_v)^2\big)(\mathbf{u}_{A,v}^{\perp T}\mathbf{\Lambda}_A \mathbf{u}_{A,v}^{\perp} + \mathbf{u}_{B,v}^{\perp T}\mathbf{u}_{B,v}^{\perp}) \\ &\quad + 2\mathbf{u}_v^T \mathbf{e}_v \sqrt{1 - (\mathbf{u}_v^T \mathbf{e}_v)^2}\mathbf{e}_{A,v}\mathbf{\Lambda}_A \mathbf{u}_{A,v}^{\perp} \\ &\xrightarrow{p} \phi(\lambda_v)^2(\lambda_v - 1) + 1. \end{aligned} \tag{30}$$

From (29) and (30),

$$\sqrt{\frac{E(q_v^2)}{E(p_{vi}^2)}} \to \sqrt{\frac{\phi(\lambda_v)^2(\lambda_v - 1) + 1}{\rho(\lambda_v)}} = \frac{(\lambda_v - 1)}{(\lambda_v + \gamma - 1)}. \tag{31}$$

6.9. *Proof of Theorem 3.* 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

$$h(x) = \sum_{v=1}^{k} \rho^{-1}(r_v x) + p - k - x. \tag{32}$$

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

$$\frac{\partial h(x)}{\partial x} = \frac{1}{2}\sum_{v=1}^{k} r_v + \frac{1}{2}\sum_{v=1}^{k} \frac{(xr_v - (1+\gamma))r_v}{\sqrt{(xr_v - (1+\gamma))^2 - 4\gamma}} - 1, \tag{33}$$

$$\frac{\partial^2 h(x)}{\partial x^2} = 2\sum_{v=1}^{k} \frac{-r_v^2 \gamma}{((xr_v - (1+\gamma))^2 - 4\gamma)^{3/2}} < 0, \tag{34}$$

so $h(x)$ is a concave function of x given r_v. From the fact that $\rho^{-1}(r_v p) > 1$ for $v \le k$, we know $h(p) > 0$. Because of the concave nature of this function, $h(x) = 0$ has a 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

$$\sum_{v=1}^{p} d_v = \frac{1}{n} \operatorname{trace}(\mathbf{Z}\mathbf{Z}^T \mathbf{\Lambda}) = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n} \lambda_i z_{ij}^2, \tag{35}$$

thus

$$E\left(\frac{\sum_{v=1}^{p} d_v}{\omega} \right) = \frac{\sum_{v=1}^{m} \lambda_v + p - k}{\omega} \to 1 \tag{36}$$

and

$$\operatorname{Var}\left(\frac{\sum_{v=1}^{p} d_v}{\omega} \right) = \frac{1}{n} \frac{\sum_{v=1}^{p} \lambda_v^2}{\omega^2} \left(E(z_{11}^4) - 1 \right) \to 0. \tag{37}$$

By (36) and (37),

$$\sum_{v=1}^{p} d_v / \omega = 1 + o_p(1). \tag{38}$$

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

$$\tilde{d}_v = d_v \omega \Big/ \sum_{v=1}^{p} d_v = d_v \left(1 + o_p(1)\right) \xrightarrow{p} \rho(\lambda_v). \tag{39}$$

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

$$h(\omega) = \sum_{v=1}^{k} \rho^{-1}(\tilde{d}_v) - \sum_{v=1}^{k} \lambda_v = o_p(1). \tag{40}$$

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

$$\omega = \tau + o_p(1). \tag{41}$$

Therefore,

$$\hat{d}_v = r_v \tau = r_v \left(\omega + o_p(1)\right) = \tilde{d}_v + o_p(1) \xrightarrow{p} \rho(\lambda_v) \tag{42}$$

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

REFERENCES

[1] AHN, J., MARRON, J. S., MULLER, K. M. and CHI, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760–766. MR2410023

[2] ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148. MR0145620

[3] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

[4] BAI, Z. and YAO, J.-F. (2008). Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. H. Poincaré Probab. Statist.* **44** 447–474. MR2451053

[5] BAI, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9** 611–677. MR1711663

[6] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575

[7] BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680

[8] BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436

[9] BOVELSTAD, H., NYGARD, S., STORVOLD, H., ALDRIN, M., BORGAN, O., FRIGESSI, A. and LINGJAERDE, O. (2007). Predicting survival from microarray data a comparative study. *Bioinformatics* **23** 2080–2087.

[10] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012

[11] FELLAY, J., SHIANNA, K., GE, D., COLOMBO, S., LEDERGERBER, B., WEALE, M., ZHANG, K., GUMBS, C., CASTAGNA, A., COSSARIZZA, A. ET AL. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* **317** 944–947.

[12] GIRSHICK, M. (1936). Principal components. *J. Amer. Statist. Assoc.* **31** 519–528.

[13] GIRSHICK, M. (1939). On the sampling theory of roots of determinantal equations. *Ann. Math. Statist.* **10** 203–224. MR0000127

[14] HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 427–444. MR2155347

[15] HORN, R. and JOHNSON, C. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

[16] JACKSON, J. (2005). *A User's Guide to Principal Components*. Wiley, New York.

[17] JOHNSTONE, I. and LU, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

[18] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[19] JOLLIFFE, I. (2002). *Principal Component Analysis*. Springer, New York. MR2036084

[20] JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454

[21] MA, S., KOSOROK, M. R. and FINE, J. P. (2006). Additive risk models for survival data with high-dimensional covariates. *Biometrics* **62** 202–210. MR2226574

[22] MARČENKO, V. and PASTUR, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics* **1** 457–483.

[23] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013

[24] PATTERSON, N., PRICE, A. and REICH, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2** e190.

[25] PAUL, D. (2005). Asymptotics of the leading sample eigenvalues for a spiked covariance model. Technical report. Available at http://anson.ucdavis.edu/~debashis/techrep/eigenlimit.pdf.

[26] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[27] PRICE, A., PATTERSON, N., PLENGE, R., WEINBLATT, M., SHADICK, N. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

[28] PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P., DALY, M. ET AL. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Gen.* **81** 559–575.

[29] RAO, N. R., MINGO, J. A., SPEICHER, R. and EDELMAN, A. (2008). Statistical eigeninference from large Wishart matrices. *Ann. Statist.* **36** 2850–2885. MR2485015

[30] WALL, M., RECHTSTEINER, A. and ROCHA, L. (2003). Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis* (D. P. Berrar, W. Dubitzky and M. Granzow, eds.) 91–109. Kluwer, Norwell, MA.

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
3101 MCGAVRAN-GREENBERG, CB 7420
CHAPEL HILL, NORTH CAROLINA 27599
USA
E-MAIL: slee@bios.unc.edu
fzou@bios.unc.edu
fwright@bios.unc.edu