Principal Component Analysis (PCA) for High Dimensional Data. PCA is dead. Long live PCA.

Fan Yang, Kjell Doksum, and Kam-Wah Tsui

ABSTRACT. Sample covariances and eigenvalues are famously inconsistent when the number d of variables is at least as large as the sample size n. However, when $d \gg n$, genomewide association studies (GWAS) that apparently are based on principal component analysis (PCA) and use sample covariances and eigenvalues are famously successful in detecting genetic signals while controlling the probability of false discoveries. To reiterate: "PCA is dead, long live PCA", or "PCA is the worst of methods, PCA is the best of methods". We outline recent work (Yang, 2013) that reconciles the worst/best dichotomy by acknowledging that PCA is indeed inconsistent for many classical statistical settings, but for settings that are natural in genomic studies, PCA produces effective methods. The dichotomy can in part be explained by how models are viewed and the goal of the study being carried out. We compare the effectiveness of three PCA methods for testing the association between covariates and a response in a framework with continuous variables. These methods are based on adjusting the data using PCA, then applying Pearson, Spearman and normal scores correlation tests.

1. Introduction

Because of the importance of the covariance matrix Σ and its eigenvalues to statistical analysis their accurate estimation is an important goal in statistics. With high-dimensional data where the dimension d of the random vector x is at least as large as the sample size n, the sample covariance matrix S may fail to be consistent. Because large data sets are becoming common, this is an important problem. A number of recent articles that address the problem of constructing consistent estimates of Σ in the $d \geq n$ case start by referring to the inconsistency of the sample covariance S. A typical example is "It is now well understood that in such a setting the standard sample covariance matrix does not provide satisfactory performance and regularization is needed". (Cai and Zhou (2012)). Other articles that start by referring to S as unsatisfactory and address the large d problem using regularization methods such as banding, tapering, thresholding, shrinking and penalization are by Wu and Pourahmadi (2003), Zou, Hastie and Tibshirani (2006), Bickel and Levina (2008a, b), EL Karoui (2008), Amini and Wainwright (2009), Cai, Zhang and Zhou

2010 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Eigenstrat, Eigensoft, GWAS, rank tests, stratification, dual principal components.

YANG, F., DOKSUM, K. and TSUI, K.W., "Principal component analysis (PCA) for high dimensional data. PCA is dead. Long live PCA. In: Ahmed, S.E. (ed.), Proceedings for Workshop on Perspectives on High Dimensional Data Analysis II, Montreal. Comtemporary Mathematics, American Mathematical Society, Providence, Rhode Island (2014).

(2009), Lam and Fan (2009), Johnstone and Lu (2009), Ahmed and Raheem (2012), Ma (2012) and Deng and Tsui (2013), among others.

On the other hand, in genomics, PCA based on sample covariance matrices and their eigenvalues have been used to construct effective tests of association between genetic marker scores and disease indicators. One collection of genomewide association studies (GWAS) based on the methodology "Eigenstrat", or its updated and expanded version "Eigensoft", started with the papers by Price *et al.* (2006) and Patterson *et al.* (2006). For a statistical examination of GWAS for case-control studies, see Lin and Zeng (2011).

The discrepancy between PCA in High Dimensional Data Analysis (HDDA) being "unsatisfactory" in statistics and "effective" in genomics can be explained by the phase "in such a setting" in the Cai and Zhou above quote. Here we examine settings where PCA is effective. In particular, we show that PCA is effective in HDDA when (i): the response vector is split into a low dimensional vector containing the responses of initial interest and a high dimensional vector of potentially confounding covariates, and (ii): the sample is drawn from a population made up of unknown subpopulations or strata and this population stratification has the potential to create confounding variables that lead to spurious correlation between a response and predictors.

Sections 2, 3 and 4 provide a summary of our framework taken from Yang (2013). Section 5 uses simulations to show and compare the effectiveness of these PCA methods.

2. Association regression models based on PCA.

2.1. Principal components. Population PCA for the random vector $\boldsymbol{x} = (X_1, \cdots, X_d)^T$ first produces a measure of the variability of $\boldsymbol{x}$ by finding the linear combination $e^T\boldsymbol{x}$ that has maximal normalized variance $Var\left(e^T\boldsymbol{x}\right)/\|e\|^2$. Let Σ denote the covariance matrix of $\boldsymbol{x}$, then $\boldsymbol{e_1}$, the first eigenvector, is

$$\boldsymbol{e_1} = \underset{e:\|e\|=1}{\operatorname{argmax}}\{e^T\Sigma e\} \tag{2.1}$$

and the first eigenvalue and the first principal component (PC_1) are

$$\lambda_1 = \boldsymbol{e_1}^T\Sigma\boldsymbol{e_1}, \;\; PC_1 = \boldsymbol{e_1}^T\boldsymbol{x}.$$

The second eigenvector $\boldsymbol{e_2}$, second eigenvalue λ_2, and second PC are obtained in the same way except $\boldsymbol{e_2}$ is found by maximizing (2.1) over e orthogonal to $\boldsymbol{e_1}$. To obtain $\boldsymbol{e_k}$, λ_k and PC_k, (2.1) is maximized over e orthogonal to $\boldsymbol{e_1}, \cdots, \boldsymbol{e_{k-1}}$. This process produces the principal components $PC_1, \cdots, PC_d$ that capture much of the variability of $\boldsymbol{x}$ in the sense that $Var(PC_j) = \lambda_j$ and $\sum_{j=1}^{d}\lambda_j = \sum_{j=1}^{d} Var(X_j)$.

2.2. Regression and association studies. Suppose Y is a response variable and that $\boldsymbol{x} \in \mathbb{R}^d$ is a random predictor. In association studies, the null hypothesis H_{0k} that Y and X_k are independent is tested for one X_k at a time. Thus what is needed is a test statistic T_k whose null distribution is known; at least asymptotically, when the null hypothesis H_{0k} holds, $k = 1, \cdots, d$. In this framework, $\boldsymbol{x}_{-k} = \{X_j : 1 \le j \le d, j \ne k\}$ are confounding variables that could lead to spurious correlation

between X_k and Y. Linear analysis based on the linear model

$$Y = \alpha_k + \beta_k X_k + \sum_{j \neq k} \beta_j X_j + \epsilon \tag{2.2}$$

does not provide stable estimates unless a sample of size $n \gg d$ is available. This has led to the introduction of shrinkage methods, penalty methods and methods based on models with sparse covariance matrices. Some of the references can be found in Section 1.

In this paper we consider the case where confounding is due to population stratification and use PCA applied to x_{-k} to correct for such stratification. In particular, the $\sum_{j \neq k} \beta_j X_j$ term in (2.2) will be replaced by a sum $\sum_{j=1}^{q} \eta_j Z_j$ where the Z_j represents principal components based on x_{-k} and $q \leq 10$. To find the Z_j, we use dual PCA, which we introduce in the next section. Under certain assumptions, these Z_j's are effective indicators of which stratum an individual belongs to.

3. Dual eigenanalysis and models for stratified populations

3.1. Stratified populations. A stratified population with K strata or subpopulations $S_1, \cdots, S_K$ is such that when one member of the population is selected, the probability that the member is from subpopulation S_k is π_k, where $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k > 0$, $1 \leq k \leq K$. Consider n independent draws and let N_k be the number of draws from S_k, then $\boldsymbol{N} = (N_1, \cdots, N_K)^T$ follows the multinomial distribution $MN(n, \pi_1, \cdots, \pi_K)$ where $\sum_{k=1}^{K} N_k = n$. This strata information is not available. Instead we have n independent draws from a population that contain K unknown strata. That is, K and $\boldsymbol{N}$ are unobservable.

Consider a random vector $(X_1, \cdots, X_d)^T$ whose covariance matrix Σ is assumed to exist. We have available a $n \times d$ random data matrix $\boldsymbol{X} = (X_{ij})_{n \times d}$ where the random vectors $\boldsymbol{x_i} = (X_{i1}, \cdots, X_{id})^T$, $1 \leq i \leq n$, are independent and identically distributed. When the $\boldsymbol{x_i}$ are drawn from a stratified population the major variability of $\boldsymbol{X} = (X_{ij})$ as we change i is due to this stratification, and this variability can be examined by considering the $n \times n$ dual covariance matrix defined by

$$\hat{\Sigma}_D = d^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{X} - \bar{\boldsymbol{X}})^T, \tag{3.1}$$

where $\boldsymbol{X} - \bar{\boldsymbol{X}}$ is the $n \times d$ matrix with entries $(X_{ij} - \bar{X}_j)$, and $\bar{X}_j = n^{-1} \sum_{i=1}^{n} X_{ij}$.

To interpret $\hat{\Sigma}_D$, let $\boldsymbol{W}^{(d)} = (W_1^{(d)}, \cdots, W_n^{(d)})^T$ be the result of one random draw from the collection of n-vectors

$$\{(X_{1j}, \cdots, X_{nj})^T : 1 \leq j \leq d\}.$$

Then

$$\hat{\Sigma}_D = Cov(\boldsymbol{W}^{(d)} - \bar{\boldsymbol{W}}^{(d)}) \equiv E[(\boldsymbol{W}^{(d)} - \bar{\boldsymbol{W}}^{(d)})(\boldsymbol{W}^{(d)} - \bar{\boldsymbol{W}}^{(d)})^T],$$

where $\bar{\boldsymbol{W}}^{(d)} = (n^{-1} \sum_{i=1}^{n} W_i^{(d)})\mathbf{1}$ and $\mathbf{1}$ is a n-vector of 1's.

It is known that $\hat{\Sigma}_D$ has the same nonzero eigenvalues, up to a constant d/n, as the usual covariance matrix

$$\hat{\Sigma} = n^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}}).$$

There is another simple relationship between PCA of $\hat{\Sigma}$ and $\hat{\Sigma}_D$: let $\hat{\lambda}_q > 0$ be the qth largest eigenvalue of $\hat{\Sigma}$, then the qth principal component of $\hat{\Sigma}$ evaluated at x_i equals the ith entry of the qth eigenvector of $\hat{\Sigma}_D$, up to a constant.

One advantage of $\hat{\Sigma}_D$ is that in HDDA its dimension $n \times n$ is much smaller than the dimension $d \times d$ of $\hat{\Sigma}$. Another advantage is that if we explicitly model stratification, then we find that even though $\hat{\Sigma}_D$ is computed without using strata information, a conditional eigenanalysis of $\hat{\Sigma}_D$ reveals the unknown population stratification and provides methods for adjusting for stratification. This is because for models that include stratification, the variability measured by $\hat{\Sigma}_D$ is mainly due to the x_i coming from different strata. To show this we have to change our model framework and evaluate strata-blind methods from the point of view of an evaluator who knows the strata information. We do this in the next subsection.

3.2. The conditional model framework. Because $X_{ij} - \bar{X}_j = X_{ij} - \mu_j + O_p(n^{-1})$, $W_1^{(d)} - \bar{W}^{(d)}, \cdots, W_n^{(d)} - \bar{W}^{(d)}$ are nearly independent, and the covariance matrix $\hat{\Sigma}_D$ would appear to be nearly diagonal. Thus eigenanalysis of $\hat{\Sigma}_D$ would appear to lack the ability to measure stratification effects. Using Jung and Marron (2009), we can prove that unconditional eigenanalysis of $\hat{\Sigma}_D$ leads to inconsistent methods. But here is where we change "the setting" of the usual statistical eigenanalysis: our methods do not use the strata information, but we evaluate the performance of the methods assuming that N_k of the $W_i^{(d)}$, $1 \le i \le n$, come from stratum k, $k = 1, \cdots, K$. Without loss of generality, we order $W_i^{(d)}$ so that $W_1^{(d)}, \cdots, W_{N_1}^{(d)}$ are from stratum 1, and so on. In this setting where we condition on the unobservable $\boldsymbol{N} = (N_1, \cdots, N_K)^T$,

$$E_{\boldsymbol{N}}(W_i^{(d)}) = \mu_k, \qquad N_k + 1 \le i \le N_{k+1}, \ 0 \le k \le K-1,$$

where μ_k is the kth stratum mean and $N_0 = 0$. It follows that in this setting, for $\mu = \sum_{k=1}^{K} \hat{\pi}_k \mu_k$, where $\hat{\pi}_k = N_k/n$,

$$E_{\boldsymbol{N}}(W_i^{(d)} - \bar{W}^{(d)}) = \mu_k - \mu, \qquad N_k + 1 \le i \le N_{k+1}, \ 0 \le k \le K-1. \tag{3.2}$$

This implies that for stratified populations, the dual covariance matrix measures the deviation of the individual strata means from the overall mean, and in this setting both the diagonal and off-diagonal elements of

$$\Sigma_D^{(\boldsymbol{N})} = E_{\boldsymbol{N}}[(\boldsymbol{W}^{(d)} - \bar{\boldsymbol{W}}^{(d)})(\boldsymbol{W}^{(d)} - \bar{\boldsymbol{W}}^{(d)})^T]$$

measure the extent of the stratification. For the ith individual in a sample of n individuals, the variables $X_{i1}, \cdots, X_{id}$ provides information about which strata the individuals is likely to belong to, and this information can be obtained from the eigenanalysis of $\hat{\Sigma}_D$.

Remark 3.1 In our conditional framework, $\Sigma_D^{(\boldsymbol{N})}$ is not a covariance matrix because $\boldsymbol{W}^{(d)} - \bar{\boldsymbol{W}}^{(d)}$ does not have mean zero (see (3.2)). This makes our analysis different from the statistical literature that reports on the failure of PCA.

3.3. Dual eigenanalysis for $d \gg n$. To examine the $d \gg n$ case, we let $d \to \infty$ and let n be finite or we let $n \to \infty$ in the $d = \infty$ universe. The elements of $\hat{\Sigma}_D$ are averages of d terms. A basic assumption is that these averages satisfy

the weak law of large numbers. That is, as $d \to \infty$, we assume that there exists $\boldsymbol{W} = (W_1, \cdots, W_n)^T$ such that with $\bar{\boldsymbol{W}} = (n^{-1} \sum_{i=1}^{n} W_i)1$,

$$\hat{\Sigma}_D \xrightarrow{P} \Sigma_D^{(N)} \equiv E_{\boldsymbol{N}}[(\boldsymbol{W} - \bar{\boldsymbol{W}})(\boldsymbol{W} - \bar{\boldsymbol{W}})^T], \tag{3.3}$$

where W_i is interpreted to be the probability limit of $W_i^{(d)}$ as $d \to \infty$ for the conditional model in Section 3.2.

From (3.2), we see that because $E_{\boldsymbol{N}}(W_i^{(d)} - \bar{W}^{(d)})$ is the same for all d,

$$E_{\boldsymbol{N}}(W_i - \bar{W}) = \mu_k - \mu, \qquad N_k + 1 \le i \le N_{k+1}, \ 0 \le k \le K-1.$$

For the ith individual in a sample of size n, we have available an infinite sequence $X_{i1}, X_{i2}, \cdots$ that provides information about which strata the individual belongs to. This information is provided by W_i. It is known (Yang (2013)) that when sampling from a stratified population with K strata, $\Sigma_D^{(N)}$ has $(K-1)$ eigenvalues of order n, one that equals zero and the rest that are constant. Moreover, the $(K-1)$ eigenvectors of $\Sigma_D^{(N)}$ corresponding to the $(K-1)$ largest eigenvalues $\lambda_1 > \cdots > \lambda_{K-1}$ capture the variability of $W_1, \cdots, W_n$ due to stratification. In genomics, Patterson *et al.* (2006) and Price *et al.* (2006) argue convincingly using real data that the number of strata is at most eleven. In the next section we will use the eigenvectors corresponding to the q largest eigenvalues of $\Sigma_D^{(N)}$, where $q \le 10$.

Remark 3.2 Dual eigenanalysis can also be accomplished using singular value decomposition. The approach taken here is more in line with the intuition provided by Section 3.3.

3.4. Classical eigenanalysis for $d \gg n$. The above is a brief summary of the PCA of $\hat{\Sigma}_D$ and $\Sigma_D^{(N)}$ for the $d \gg n$ case from Yang (2013). Next we turn to a summary of HDDA PCA for $\hat{\Sigma}$ and Σ. It is known that the m largest eigenvalues $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_m > 0$ of $\hat{\Sigma}$ and $\hat{\Sigma}_D$ are the same. Let $\widehat{\boldsymbol{PC}}_q$ be the n-vector obtained by evaluating the qth PC of $\hat{\Sigma}$ at $\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}$, and let $\hat{\boldsymbol{e}}_{\boldsymbol{Dq}}$ be the eigenvector of $\hat{\Sigma}_D$ corresponding to $\hat{\lambda}_q$, then (Yang(2013))

$$\hat{\lambda}_q^{-\frac{1}{2}} \widehat{\boldsymbol{PC}}_q = \hat{\boldsymbol{e}}_{\boldsymbol{Dq}}, \quad q = 1, \cdots, m. \tag{3.4}$$

That is, the PC's of $\hat{\Sigma}$ are equivalent to the eigenvectors of $\hat{\Sigma}_D$. Because (3.4) holds for all d, we represent the $d \gg n$ PC's of $\hat{\Sigma}$ for the conditional model as the equivalence class $\{c\boldsymbol{e}_{Dq}^{(N)} : c > 0\}$, where $\boldsymbol{e}_{Dq}^{(N)}$ is the qth eigenvector of $\Sigma_D^{(N)}$. In practice any c will work. We set $c = 1$ and represent the $d \gg n$ PC of $\hat{\Sigma}$ as

$$\boldsymbol{PC_q}^{(N)} = \boldsymbol{e_{Dq}}^{(N)}.$$

Simulation results (Price *et al.* (2006), Yang (2013)) have shown that for stratified populations, when $d \gg n$, creating a sparse model where $\boldsymbol{x}_{-k}$ is replaced by $\boldsymbol{e}_{Dj}^{(N)}$, $1 \le j \le 10$, controls for the confounding effects of $\boldsymbol{x}_{-k}$ on the correlation between Y and X_k.

3.5. Regression models for stratified populations. Based on the previous discussion, we consider sparse models based on the first q PC's based on the usual sample covariance matrix $\hat{\Sigma}$, where $q \le 10$. To simplify notation, in this subsection, let X be the X_k whose association with Y is being examined and let

$\boldsymbol{X} \in \mathbb{R}^{d-1}$ be the vector of confounding predictors. The discussion in Section 3.3 shows that a reasonable sparse model is

$$Y_i = \alpha + \beta X_i + \sum_{j=1}^{q} \eta_j Z_{ij} + \epsilon_i, \quad 1 \le i \le n \tag{3.5}$$

where Z_{ij} is the $d = \infty$ version of the ith entry of the jth dual population eigenvector based on $\boldsymbol{X}$. Extending results of Patterson *et al.* (2006), Yang (2013) has shown that when $K < q$, Z_{ij} is an indicator of which strata j the ith observation comes from for $j = 1, \cdots, K$; while Z_{ij} is "noise" for $j = K+1, \cdots, q$. Assume a trivariate normal distribution, or more generally assume that,

$$Y_i = a + b^T \boldsymbol{Z_i} + \epsilon_{1i},$$
$$X_i = c + d^T \boldsymbol{Z_i} + \epsilon_{2i},$$

where $\boldsymbol{Z_i} = (Z_{i1}, \cdots, Z_{iq})^T$. Then we can rewrite this model as (see Bickel and Doksum (2007), Example 6.2.1)

$$Y_i - Y_{i0} = \alpha + \beta(X_i - X_{i0}) + \delta_i, \tag{3.6}$$

where the δ_i have mean zero and the same variance τ^2, and X_{i0} and Y_{i0} are the linear predictors of X_i and Y_i based on $\boldsymbol{Z}$ that minimize the mean squared prediction error.

4. Association test statistics

When the variables X_i and δ_i in model (3.6) are Gaussian, the test statistic based on the MLE of β is a 1-1 function of the Pearson correlation coefficient of the strata adjusted variables

$$X^A = X - X_0, \qquad Y^A = Y - Y_0.$$

Here X_0 and Y_0 are unknown and will be replaced by their predictors

$$\hat{X}_0 = \bar{X} + \left(\widehat{Cov}^{-1}(\hat{\boldsymbol{Z}})\widehat{Cov}(\hat{\boldsymbol{Z}}, X)\right)^T (\hat{\boldsymbol{Z}} - \bar{\hat{\boldsymbol{Z}}}), \tag{4.1}$$

$$\hat{Y}_0 = \bar{Y} + \left(\widehat{Cov}^{-1}(\hat{\boldsymbol{Z}})\widehat{Cov}(\hat{\boldsymbol{Z}}, Y)\right)^T (\hat{\boldsymbol{Z}} - \bar{\hat{\boldsymbol{Z}}}), \tag{4.2}$$

where $\widehat{Cov}$ refers to sample covariance matrix and $\hat{\boldsymbol{Z}}$ is the $q \times 1$ vector of sample PC's. Our test statistics will be based on the stratum adjusted variables

$$\hat{X}^A = X - \hat{X}_0, \ \hat{Y}^A = Y - \hat{Y}_0. \tag{4.3}$$

For a sample $\{(x_i, Y_i),\ 1 \le i \le n\}$, we compute $\{(\hat{X}_{ij}^A, \hat{Y}_i^A),\ 1 \le i \le n\}$ according to (4.3) and define

$$R_P = \text{Sample corr}\{(\hat{X}_{ij}^A, \hat{Y}_i^A),\ 1 \le i \le n\}.$$

This is the Pearson's type correlation coefficient. Because the adjusted X and Y are not Gaussian, it makes sense to consider robust statistics such as the Spearman and the normal scores rank statistics

$$R_S = \text{Sample corr}\{(Q_i, R_i) :\ 1 \le i \le n\},$$
$$R_{NS} = \text{Sample corr}\{(\Phi^{-1}(\frac{Q_i}{n+1}), \Phi^{-1}(\frac{R_i}{n+1})) :\ 1 \le i \le n\},$$

where Q_i and R_i are the ranks of $\hat{X}_{ij}^A$ and $\hat{Y}_i^A$ among $(\hat{X}_{1j}^A, \cdots, \hat{X}_{nj}^A)$ and $(\hat{Y}_1^A, \cdots, \hat{Y}_n^A)$, respectively, and Φ is the cumulative distribution function of $N(0,1)$. All these statistics have the asymptotic standard normal distribution $N(0,1)$ under the null hypothesis $H_0 : \beta = 0$. Thus we use standard normal critical values for our hypothesis testing.

5. Simulations

In this section we compare the Type I error probabilities and power of the three methods of Section 4 for testing the association between a predictor X_k and a response Y when sampling from a population with $K = 2$ strata. The strata information is unobservable, but researchers typically use methods based on $q = K - 1 = 10$. Thus we also compare properties of the methods based on $q = 1$ and $q = 10$ to examine whether using q too large leads to a loss in performance. Another question is what are the properties of the tests if stratification is ignored, that is, the statistician uses $q = 0$. It is found that in this case the Type I error probabilities are far off target, so a power comparison when $q = 0$ is not of interest. Section 4 type tests with $q = 0$ should not be used for stratified populations.

We generate a $n \times d$ data matrix $\boldsymbol{X}_{n\times d}$ for a population with two strata column by column. Let n_1 and n_2 be positive integers such that $n_1 + n_2 = n$. For $j = 1, \cdots, d$, we generate the jth column as follows. We first fix the value of ϕ and then generate μ_1 and μ_2 i.i.d. from $N(0, \phi^2)$. Then n_1 independent values for sample members are generated from the first stratum as $N(\mu_1, 1)$, and independently, n_2 independent values for sample members are generated from stratum 2 as $N(\mu_2, 1)$. The jth column consists of these n_1 and n_2 generated values. The d columns are generated independently. For our simulation experiment, we set $n_1 = 600$, $n_2 = 400$, (hence $n = n_1 + n_2 = 1000$) and ϕ is set as 1, 2 or 3.

Next we generate $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ using the model

$$Y_i = \alpha_i + \beta X_{ik} + \eta Z_{i1} + \epsilon_i, \quad 1 \le i \le n \tag{5.1}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.75, 1$ or 1.25, X_{ik} is the ith observation for the variable X_k that is generated as in the previous paragraph, and Z_{i1} is the ith entry of $\boldsymbol{Z}_1$.

In order to compute the strata adjusted values (4.3), which the test statistics are based on, we compute the normalized sample PC vectors using simulated data, denoted by $\hat{\boldsymbol{Z}}_1 = \widehat{\boldsymbol{PC}}_1/\|\widehat{\boldsymbol{PC}}_1\|, \cdots, \hat{\boldsymbol{Z}}_q = \widehat{\boldsymbol{PC}}_q/\|\widehat{\boldsymbol{PC}}_q\|$. The model is fixed with the true number of strata $K = 2$, but the statistician does not know K and checks the performance of his/her methods using $q = 0$ (ignoring the stratification), $q = 1$ (using the correct number of PC's) or $q = 10$ (being overly conservative).

To investigate Type I error, we generate data using model (5.1) with $\alpha = 1$, $\beta = 0$, $\eta = 1$, $n = 1{,}000$ and $d = 10{,}000$. Repeat the above experiment for 50 times and compute the proportion

$$a = \frac{\#\text{ Rejections}}{10{,}000 \times 50}$$

as the estimate of the probability of Type I error.

To examine Power $= 1 - P$ (Type II error), we generate data sets with $n = 1{,}000$, $10{,}000$ irrelevant variables ($\beta = 0$) and 200 relevant variables using model (5.1) with $\alpha = 1$, $\beta = 0.1$ and $\eta = 1$. We then have $d = 10{,}200$ variables. Repeat

the above experiment for 50 times. Then the estimated power is

$$b = \frac{\#\text{ Rejections among the 200 relevant variables}}{200 \times 50}.$$

Table 1 shows that without the adjustment for stratification, the three tests fail to achieve the correct significance level, while all three succeed when using one or ten PC's to correct for stratification.

TABLE 1. Monte Carlo Type I error probabilities for Gaussian $N(0,1)$ errors. The nominal significance level is 10^{-3}.

	$\phi = 1$	$\phi = 2$	$\phi = 3$
Pearson 0 PC	2.92×10^{-3}	5.04×10^{-3}	6.39×10^{-3}
Spearman 0 PC	2.81×10^{-3}	4.49×10^{-3}	5.10×10^{-3}
Normal Scores 0 PC	2.86×10^{-3}	4.22×10^{-3}	4.85×10^{-3}
Pearson 1 PC	1.08×10^{-3}	1.04×10^{-3}	1.04×10^{-3}
Spearman 1 PC	1.05×10^{-3}	1.06×10^{-3}	1.05×10^{-3}
Normal Scores 1 PC	1.04×10^{-3}	1.04×10^{-3}	1.04×10^{-3}
Pearson 10 PC	1.14×10^{-3}	1.14×10^{-3}	1.15×10^{-3}
Spearman 10 PC	1.07×10^{-3}	1.10×10^{-3}	1.10×10^{-3}
Normal Scores 10 PC	1.14×10^{-3}	1.14×10^{-3}	1.14×10^{-3}

TABLE 2. Power for Gaussian $N(0, \sigma_\epsilon^2)$ errors. The significance level is 10^{-3}.

	$\sigma_\epsilon = 0.75$	$\sigma_\epsilon = 1$	$\sigma_\epsilon = 1.25$
Pearson 1 PC	0.8175	0.4390	0.2171
Spearman 1 PC	0.7597	0.3792	0.1851
Normal Scores 1 PC	0.8154	0.4366	0.2164
Pearson 10 PC	0.8156	0.4391	0.2175
Spearman 10 PC	0.7576	0.3821	0.1847
Normal Scores 10 PC	0.8130	0.4358	0.2150

TABLE 3. Power for $\gamma N(0,1) + (1-\gamma)Unif(-2,2)$ errors. The significance level is 10^{-3}.

	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$
Pearson 1 PC	0.2833	0.3618	0.3968	0.4281
Spearman 1 PC	0.2420	0.3153	0.3466	0.3769
Normal Scores 1 PC	0.6530	0.5519	0.5018	0.4642
Pearson 10 PC	0.2850	0.3637	0.3967	0.4273
Spearman 10 PC	0.2338	0.3085	0.3419	0.3714
Normal Scores 10 PC	0.5825	0.5141	0.4780	0.4561

Table 2 shows that the Pearson and normal scores tests perform equally well and better than the Spearman test for Gaussian errors. Moreover, very little power is lost by using 10 PC's instead of one, which is the "true" number of PC's.

We also tried contaminated distributions for the error ϵ and found that for $0.9N(0,1) + 0.1N(0,4)$ and $0.9N(0,1) + 0.1N(10,9)$, the results are very similar to Tables 1 and 2. However, for the model where the distribution of ϵ has the form $\gamma N(0,1) + (1-\gamma)Unif(-2,2)$ with $\gamma \leq 0.9$, the normal scores test is best, Pearson is second best, and Spearman has the lowest power. See Table 3. (The Type I errors are still controlled at the significance level for the non-Gaussian errors (not shown).) All the tables show that when the true number of strata is 2, using methods that assume 11 strata ($q = 10$) do not lose much in performance when compared to methods using the correct $q = 1$.

References

1. Ahmed, S.E. and Raheem, S.E. (2012). Shrinkage and absolute penalty estimation in linear models. *WIREs Computational Statistics* **4**, 541-553.
2. Amini, A.A. and Wainwright, M.J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal commponents. *Ann. Statist.* **37**, 2877-2921.
3. Bickel, P.J. and Doksum, K.A. (2007). *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I.* Updated Printing. Pearson, Upper Saddle River, New Jersey.
4. Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.
5. Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.
6. Cai, T.T. and Zhou H.H. (2012). Minimax estimation of large covarinace matrices under l_1-norm. *Statistica Sinica* **22**, 1319-1378.
7. Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38**, 2118-2144.
8. Deng, X. and Tsui, K.W. (2012). Penalized covariance estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics*, in press.
9. D'Aspremont, A., Ghaoui, L.EI, Jordan M., and Lanckriet G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49**, 434-448.
10. El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717-2756.
11. Johnstone, I.M. and Lu, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Ann. Statist.* **37**, 4254-4278.
12. Jolliffe, I., Trendafilov, N. and Uddin M. (2003). A modified principal component techique based on the LASSO. *Journal of American Statistical Association* **104**, 682-693.
13. Jung, S. and Marron J.S. (2009). PCA Consistency in High Dimension, Low Smaple Size Context. *Ann. Statist.* **37**, 4104-4130.
14. Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Ann. Statist.* **37**, 4254-4278.
15. Lin, D.Y. and Zeng, D. (2011). Correcting for population stratification in genomewide association studies. *Journal of the Aemerican Statistical Association* **106:495**, 997-1008.
16. Ma, Z. (2012). Accuracy of the Tracy-Widom limits for the extreme eigenvalues in white Wishart matrices. *Bernoulli* **18**, 322-359.
17. Patterson, N., Price, A. and Reich, D. (2006). Population structure and Analysis. *PLoS Genet* **2**, e190.
18. 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.
19. Shen, H. and Huang J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99**, 1015-1034.
20. Ulfarsson M.O. and Solo V. (2008). Sparse variable PCA using geodesic steepest descent. *IEEE T. Signal Proces.* **56**, 5823-5832.
21. Witten, D.M., Tibshirani R. and Hastie T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515-534.

22. Wu, W.B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831-844.
23. Yang, F. (2013). On high dimensional data analysis and biomedical genomics. PhD Thesis, University of Wisconsin, Madison.
24. Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265-286.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706
E-mail address: fyang@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706
E-mail address: doksum@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706
E-mail address: kwtsui@stat.wisc.edu