Adjusting systematic bias in high dimensional principal component scores

Sungkyu Jung

Seoul National University

Abstract: Principal component analysis continues to be a powerful tool in dimension reduction of high dimensional data. We assume a variance-diverging model and use the high-dimension, low-sample-size asymptotics to show that even though the principal component directions are not consistent, the sample and prediction principal component scores can be useful in revealing the population structure. We further show that these scores are biased, and the bias is asymptotically decomposed into rotation and scaling parts. We propose methods of bias-adjustment that are shown to be consistent and work well in the high dimensional situations with small sample sizes. The potential advantage of bias-adjustment is demonstrated in a classification setting.

Key words and phrases: HDLSS, jackknife, principal component analysis.

1. Introduction

Principal component analysis (PCA) is a workhorse method of multivariate analysis, and has been used in a variety of fields for dimension

reduction, visualization and as exploratory analysis. The standard estimates of principal components, obtained by either the eigendecomposition of the sample covariance matrix or the singular value decomposition of the data matrix, are now well-known to be inconsistent when the number of variables, or the dimension d, is much larger than the sample size n (Paul, 2007; Johnstone & Lu, 2009; Jung & Marron, 2009). These observations were paralleled with a vast amount of proposals on, e.g., sparse principal component estimations (*cf.* most notably, Zou et al., 2006), which perform better in some models with high dimensions.

However, the standard estimates of principal components (PCs) continue to be useful, partly due to fast computations available (see, e.g., Abraham & Inouye, 2014). Many of the sparse estimation methods, unfortunately, do not computationally scale well for large data with hundreds of thousands of variables. Moreover, the standard estimation has shown to be useful in some application areas such as imaging, genomics and big-data analysis (Fan et al., 2014). In these areas, the sample and prediction PC scores (the projection scores of the data points onto the PC directions) are often used in the next stage of analysis.

The prediction of PC scores has considerable practical utility in modern data analysis. A prominent example where the "sample" and "prediction"

PC scores are used is the *PC regression.* In particular, for prediction and cross-validation for PC regression, the PC scores are used as explanatory variables. For prediction of the response from a new set of observations, the predicted PC scores are needed (Jackson, 2005). As an example, Li et al. (2014) used a PC regression in prediction of an phytoplankton abundance index. In the same vein, *classification* rules are often estimated for dimension-reduced data sets. As an instance, in forensic science, residue features from various black ballpoint inks are dimension-reduced (via PCA) then classified, based on a lab data set. New features from the field are classified using their prediction scores as an input for the classification rule (Adam et al., 2008). As a more involved example, ancestry estimation in genetic association studies uses the sample PC scores obtained from a reference genotyped sample, often from large-scale public sequencing data sets (Zhan et al., 2013; Marcus et al., 2020; Wang et al., 2015). The prediction PC scores of a new sample is then matched to the sample PC scores, in order to infer the new samples's ancestry membership (Zhang et al., 2020).

In this paper, we revisit the standard estimates of principal components in ultra-high dimensions and reveal that while the component directions and variances are inconsistent, the sample and prediction scores are useful for moderately large sample size. For low sample sizes, the scores are biased.

We quantify the bias, decompose it into two systematic parts, and propose to estimate bias-adjustment factors.

As a visual example of the systematic bias, a toy data set with 2 distinguishable principal components is simulated and plotted in Fig. 1. Each observation in the data set consists of $d = 10,000$ variables. The first two sample principal component directions are estimated from $n = 50$ observations, and are used to obtain the sample and prediction scores (the latter are computed from 20 new observations). The true principal scores are also plotted and connected to their empirical counterparts. This example visually reveals that the sample scores are systematically biased, that is, *uniformly rotated* and *stretched*. What is more surprising is that the prediction scores are also uniformly rotated, by the same angle as the sample scores, and uniformly shrunk.

On the other hand, the third component scores from this example appear to be quite arbitrary; see Fig. 2. (The estimates for component 3 in this example is only as good as random guess.) Moreover, unlike the first two components plotted in Fig. 1, the sample scores of the third component are grossly inflated, while the prediction scores are much smaller than the sample scores.

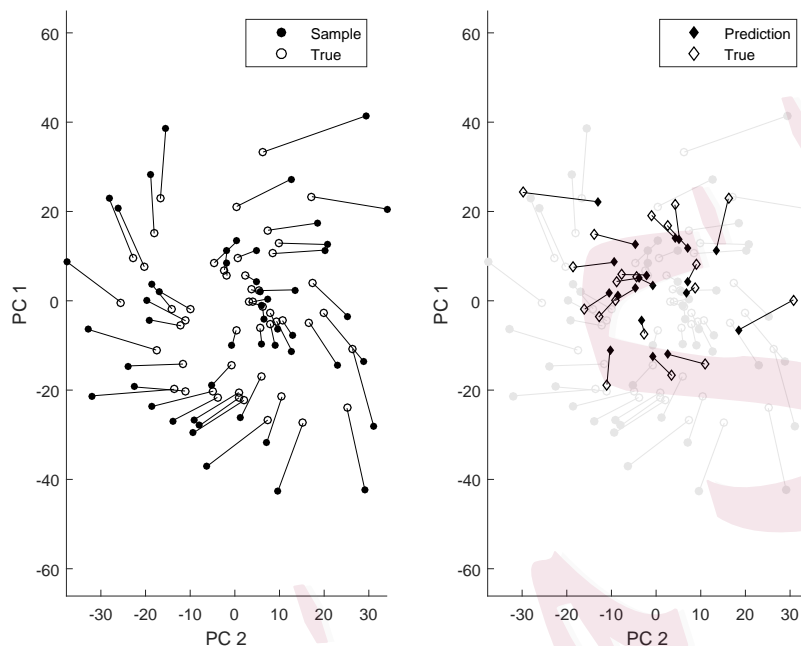In Section 2, we provide theoretical justification of the phenomenon

Figure 1: Sample and prediction principal component scores connected to their true values. This toy data set of size $(d, n) = (10000, 50)$ is generated from the spike model with $m = 2$ spikes, with polynomially-decreasing eigenvalues with $\beta = 0.3$; see Section 4.2 for details.

observed in Figs. 1 and 2, and asymptotically quantify the two parts of the systematic bias. We assume m-component models with diverging variances, and use the high-dimension, low-sample-size asymptotic scenario (i.e. $d \to \infty$ while n is fixed). These models and asymptotics are used in giving the contrasting results of the sample and prediction scores. The correlation
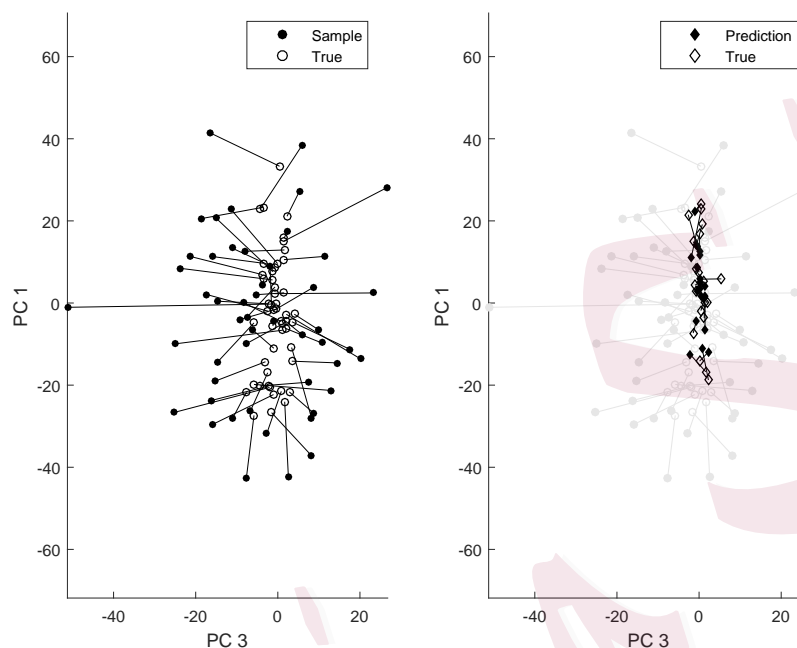
Figure 2: Sample and prediction principal component scores connected to their true values. Models and data are the same as in Fig. 1.

coefficients between the sample (or prediction) and true scores turn out to be close to 1, for large signals and large sample sizes, indicating the situations where the principal component scores are most useful.

Since the bias is asymptotically quantified, the natural next step is to adjust the bias by estimating the bias-adjustment factor. In Section 3, we propose a simple, yet consistent, estimator and several variants of estimators based on the idea of Jackknife. Adjusting these biases improves

the performance of prediction modeling, and we demonstrate its potential by an example involving classification. Results from numerical studies are summarized in Section 4.

There are several related works on the principal component scores in high dimensions (Lee et al., 2010; Fan et al., 2013; Lee et al., 2014; Sundberg & Feldmann, 2016; Shen et al., 2016; Hellton & Thoresen, 2017; Wang & Fan, 2017; Jung et al., 2018). This paper is built upon these previous findings. In particular, this paper is a continuation of the author's previous work (Jung et al., 2018), and intermediate results are borrowed from there. While the scaling and rotation of the sample scores were previously identified in Jung et al. (2018) as well as in Hellton & Thoresen (2017), the main contributions of this paper are *i)* the quantification of the asymptotic bias for the *prediction* scores, which has not been addressed, and *ii)* a consistent estimation of the bias-adjustment factor. Under the "random-matrix" asymptotic scenario, i.e., $d/n \to c \in (0, \infty)$, Lee et al. (2010) discussed a bias adjustment of principal component scores. Our work extends Lee et al. (2010) to the high-dimension, low-sample-size asymptotic scenario. Note that the asymptotic *rotational* bias was not identified in Lee et al. (2010), due to larger sample size $n \asymp d$ considered there. A survey of high-dimension, low-sample-size asymptotics can be found in Aoshima et al.

(2018).

2. Asymptotic behavior of principal component scores

2.1 Model and assumptions

Let $\mathcal{X} = [X_1, \ldots, X_n]$ be a $d \times n$ data matrix, where each X_i is mutually independent and has zero mean and covariance matrix Σ_d. Population principal components are obtained by the eigendecomposition of $\Sigma_d = U\Lambda U^{\mathrm{T}}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is the diagonal matrix of principal component variances and $U = [u_1, \ldots, u_d]$ consists of principal component directions. For a fixed m, we assume an m-component model, where the first m component variances are distinguishably larger than the rest. Specifically, the larger variances increase at the same rate as the dimension d, i.e. $\lambda_i \asymp d$, which was previously noted as the "boundary situation" (Jung et al., 2012). This diverging-variance condition seems to be more realistic than the other simpler cases $\lambda_i \gg d$ (i.e., $\lambda_i/d \to \infty$) and $\lambda_i \ll d$ (Hellton & Thoresen, 2017; Shen et al., 2016), and is satisfied for high-dimensional models used in factor analysis (Fan et al., 2013; Li et al., 2017; Sundberg & Feldmann, 2016). In a more general asymptotic scenario of $d/n \to \infty$, our condition, $\lambda_i \asymp d$, is akin to the condition, $\lim_{n\to\infty} d/(n\lambda_i) = c_i \in (0, \infty)$, assumed in Shen et al. (2016) and Wang & Fan (2017). In particular, in the *ultra-high*

dimensional case of $n \asymp \log(d)$, as defined in Fan & Lv (2008), we have $d^{1-\epsilon} \ll d/n \ll d^{1+\epsilon}$ for any $\epsilon > 0$. Thus, although not identical, the assumption $\lambda_i \asymp d/n$ of Shen et al. (2016) and Wang & Fan (2017) is similar to (A1) below, $\lambda_i \asymp d$, in the ultra-high dimensional case.

We assume that the population principal component variances satisfy the following:

(A1) $\lambda_i = \sigma_i^2 d, \ i = 1, \ldots, m, \ \sigma_1^2 \geqslant \cdots \geqslant \sigma_m^2$.

(A2) $\lim_{d \to \infty} \sum_{i=m+1}^{d} \lambda_i / d := \tau^2 \in (0, \infty)$.

(A3) There exists $B < \infty$ such that for all $i > m$, $\limsup_{d \to \infty} \lambda_i < B$.

The conditions (A2) and (A3) are used to allow λ_i for $i > m$ increase as d increases. All of our results hold when the condition (A3) is relaxed to, e.g., allow the situation that $\lambda_i \asymp d^\alpha, \ \alpha < 1/2$. Such generalization is straightforward, but invites nonintuitive technicality (see, e.g., Jung et al., 2012, 2018). By decomposing each independent observation into the first m components and the remaining term, we write

$$X_j = \sum_{i=1}^{m} \lambda_i^{1/2} u_i z_{ij} + \sum_{i=m+1}^{d} \lambda_i^{1/2} u_i z_{ij}, \quad (j = 1, \ldots, n), \tag{2.1}$$

where z_{ij} is the normalized principal component score.

(A4) For each $j = 1, 2, \ldots$, $(z_{1j}, z_{2j}, \ldots)$ is a sequence of independent random variables such that for any i, $\mathrm{E}(z_{ij}) = 0$, $\mathrm{Var}(z_{ij}) = 1$, and that

the fourth moment of z_{ij} is uniformly bounded.

2.2   Sample and prediction principal component scores

Suppose we have a data matrix $\mathcal{X} = [X_1, \ldots, X_n]$ and a vector X_*, independently drawn from the same population with principal component directions u_i. The principal component analysis is performed for data $\mathcal{X}$ and is used to predict the principal component scores of X_*.

We define the ith *true principal component scores* of $\mathcal{X}$ as the vector of n projection scores:

$$w_i^{\mathrm{T}} = u_i^{\mathrm{T}} \mathcal{X} = (w_{i1}, \ldots, w_{in}), \quad (i = 1, \ldots, d), \tag{2.2}$$

where $w_{ij} = u_i^{\mathrm{T}} X_j = \sqrt{\lambda_i} z_{ij}$. The last equality is given by the decomposition of X_j in (2.1). Likewise, the true ith principal component score of X_* is $w_{i*} = u_i^{\mathrm{T}} X_* = \sqrt{\lambda_i} z_{i*}$.

The classical estimators of the pair of the ith principal component direction and variance are $(\hat{u}_i, \hat{\lambda}_i)$, obtained by either the eigendecomposition of the sample covariance matrix $S_d = n^{-1} \mathcal{X} \mathcal{X}^T$,

$$S_d = \sum_{i=1}^{n} \hat{\lambda}_i \hat{u}_i \hat{u}_i^{\mathrm{T}},$$

or by the singular value decomposition of the data matrix,

$$\mathcal{X} = \sqrt{n} \sum_{i=1}^{n} \sqrt{\hat{\lambda}_i} \hat{u}_i \hat{v}_i^{\mathrm{T}}, \tag{2.3}$$

where $\hat{v}_i$ is the right singular vector of $\mathcal{X}$. By replacing u_i in (2.2) with its estimator $\hat{u}_i$, we define the ith *sample principal component scores* of $\mathcal{X}$ as

$$\hat{w}_i^{\mathrm{T}} = \hat{u}_i^{\mathrm{T}} \mathcal{X} = (\hat{w}_{i1}, \ldots, \hat{w}_{in}), \quad (i = 1, \ldots, n). \tag{2.4}$$

The sample principal component scores are in fact weighted right singular vectors of $\mathcal{X}$; comparing to (2.3), $\hat{w}_i = \sqrt{n\hat{\lambda}_i}\hat{v}_i$.

For an independent observation X_*, the definition (2.4) gives

$$\hat{w}_{i*} = \hat{u}_i^{\mathrm{T}} X_*,$$

which is called the ith *prediction principal component score* for X_*.

2.3   Main results

Denote $W_1 = (\sigma_i z_{ij})_{i,j} = (d^{-1/2} w_{ij})_{i,j} = d^{-1/2}[u_1, \ldots, u_m]^{\mathrm{T}} \mathcal{X}$ for the $m \times n$ matrix of the scaled true scores for the first m principal components. The ith row of W_1 is $d^{-1/2} w_i^{\mathrm{T}}$. Similarly, the scaled sample scores for the first m principal components are denoted by $\widehat{W}_1 = d^{-1/2}[\hat{u}_1, \ldots, \hat{u}_m]^{\mathrm{T}} \mathcal{X}$.

For a new observation X_*, write $W_* = d^{-1/2}(w_{1*}, \ldots, w_{m*})^{\mathrm{T}}$ and $\widehat{W}_* = d^{-1/2}(\hat{w}_{1*}, \ldots, \hat{w}_{m*})^{\mathrm{T}}$ for the scaled true scores and prediction scores, respectively, of the first m principal components.

Write $\mathcal{W} = W_1 W_1^{\mathrm{T}}$ for the scaled $m \times m$ sample covariance matrix of the first m scores. Let $\{\lambda_i(S), v_i(S)\}$ denote the ith largest eigenvalue-

eigenvector pair of a non-negative definite matrix S and $v_{ij}(S)$ denote the

jth loading of the vector $v_i(S)$. For a sequence A_d of random matrices,

we say $A_d = O_p(b_d)$ if all elements of A_d/b_d are uniformly stochastically

bounded. Note that $A_d = O_p(1)$ implies $\|A_d\|_F = O_p(1)$.

Theorem 1. *Assume the m-component model under Conditions (A1)–(A4)*

and let $n > m \geqslant 0$ be fixed and $d \to \infty$. Then, the first m sample and

prediction scores are systematically biased:

$$\widehat{W}_1 = SR^{\mathrm{T}}W_1 + O_p(d^{-1/4}), \tag{2.5}$$

$$\widehat{W}_* = S^{-1}R^{\mathrm{T}}W_* + O_p(d^{-1/2}), \tag{2.6}$$

where $R = [v_1(\mathcal{W}), \ldots, v_m(\mathcal{W})]$, $S = diag(\rho_1, \ldots, \rho_m)$, and $\rho_k = \sqrt{1 + \tau^2/\lambda_k(\mathcal{W})}$.

Moreover, for $k > m$,

$$\hat{w}_{kj} = O_p(d^{1/2}), \quad j = 1, \ldots, n, \tag{2.7}$$

$$\hat{w}_{k*} = O_p(1). \tag{2.8}$$

Our main results show that the first m sample and prediction scores are

comparable to the true scores. The asymptotic relation tells that for large d,

the first m sample scores in $\widehat{W}_1$ converge to the true scores in W_1, uniformly

rotated and scaled for all data points. It is thus valid to use the first m

sample principal scores for exploration of important data structures, and to

reduce the dimension of the data space from d to m in the high-dimension, low-sample-size context.

Theorem 1 explains and quantifies the two parts of the bias, exemplified in Fig. 1. In particular, the same rotational bias applies to both sample and prediction scores. The scaling bias factors ρ_k in the matrix S are all greater than 1. Thus, while the sample scores are all stretched, the prediction scores are all shrunk. The second part of the theorem shows that the magnitude of inflation for the sample scores of the "noise" component (see, e.g., component 3 scores in Fig. 2) is of order $d^{1/2}$. On the other hand, the prediction scores of the noise component do not diverge.

Remark 1. Suppose $m = 1$ in Theorem 1. Then the sample and prediction scores are simply proportionally-biased in the limit: $\hat{w}_{1j}/w_{1j} \rightarrow \rho_1$ and $\hat{w}_{1*}/w_{1*} \rightarrow \rho_1^{-1}$ in probability as $d \rightarrow \infty$.

Remark 2. Suppose that the limit $n \rightarrow \infty$ is taken for the expression (2.5) and (2.6). Then from the classical asymptotic results on the $m \times m$ covariance matrix $\mathcal{W}$ (*cf.* Anderson, 1963), $S = I_m + O_p(\frac{1}{n})$ and $R = I_m + O_p(\frac{1}{n})$. That is, in the limit $d \rightarrow \infty$, the limiting bias is of order n^{-1}.

The proof of Theorem 1 relies on the asymptotic behavior of the principal component direction and variance, which is now well-understood; see Jung et al. (2018) for the asymptotic regime of $d \rightarrow \infty$, n fixed; Shen et al.

(2016) and Wang & Fan (2017) for the asymptotic regime of $d \to \infty$, $n \to \infty$ and $d/n \to \infty$. For reference we restate it here.

Lemma 1. *[Theorem S2.1, Jung et al. (2018)] Assume the conditions of Theorem 1. (i) the sample principal component variances converge in probability as $d \to \infty$;*

$$d^{-1}n\hat{\lambda}_i = \begin{cases} \lambda_i(\mathcal{W}) + \tau^2 + O_p(d^{-1/2}), & i = 1, \ldots, m; \\ \tau^2 + O_p(d^{-1/2}), & i = m+1, \ldots, n. \end{cases}$$

(ii) The inner product between sample and population PC directions converges in probability as $d \to \infty$;

$$\hat{u}_i^{\mathrm{T}} u_j = \begin{cases} \rho_i^{-1} v_{ij}(\mathcal{W}) + O_p(d^{-1/2}), & i, j = 1, \ldots, m; \\ O_p(d^{-1/2}), & otherwise. \end{cases}$$

This result is abridged later in Section 2.4 for discussion. To handle prediction scores, we need in addition the following observation, summarized in Lemma 2. For each $k = 1, \ldots, m$, the kth projection score $\hat{w}_{k*}$ is decomposed into

$$\hat{w}_{k*} = \hat{u}_k^{\mathrm{T}} X_* = \sum_{i=1}^{m} w_{i*} \hat{u}_k^{\mathrm{T}} u_i + \epsilon_{k*}, \tag{2.9}$$

where $\epsilon_{k*} = \sum_{i=m+1}^{d} w_{i*} \hat{u}_k^{\mathrm{T}} u_i$. In the next lemma, we show that the "error term," ϵ_{k*}, is stochastically bounded.

Lemma 2. *Assume the m-component model with (A1)–(A4) and let $n >$ $m \geqslant 0$ be fixed. For $k = 1, \ldots, n$, $E(\epsilon_{k*}|W_1) = 0$, and*

$$\lim_{d \to \infty} Var(\epsilon_{k*} \mid W_1) = v_O^2/(\lambda_k(\mathcal{W}) + \tau^2), \quad for \ k \leqslant m; \quad (2.10)$$

$$\lim_{d \to \infty} \frac{1}{n-m} \sum_{k=m+1}^{n} Var(\epsilon_{k*} \mid W_1) = v_O^2/\tau^2, \quad (2.11)$$

where $v_O^2 = \lim_{d \to \infty} d^{-1} \sum_{i=m+1}^{d} \lambda_i^2$. As $d \to \infty$, $\epsilon_{k*} = O_p(1)$.

Lemmas 1 and 2 facilitate an interpretation of the results in Theorem 1. Intuitively, the overestimation of the sample principal variances, in Lemma 1(i), causes the sample scores to be stretched, while the inconsistency of $\hat{u}_i$ leads to smaller $\hat{u}_i^{\mathrm{T}} u_i$ in Lemma 1(ii), which then results in the deflation of the projection scores (2.9). Proofs of Theorem 1 and all other results can be found in the supplementary material.

Next result shows that the sample and true scores (or prediction and true scores) are highly correlated with each other. For this, we compute the inner product between the standardized sample scores $\hat{w}_k/\sqrt{\hat{w}_k^{\mathrm{T}} \hat{w}_k}$ and true scores $w_k/\sqrt{w_k^{\mathrm{T}} w_k}$. Define for a pair (x, y) of n-vectors $r(x, y) = x^{\mathrm{T}} y/\sqrt{x^{\mathrm{T}} x \cdot y^{\mathrm{T}} y}$, which is an empirical correlation coefficient between x and y when the mean is assumed to be zero.

Theorem 2. *Let $\zeta_{kj} = \lambda_k(\mathcal{W})/(\sum_{\ell=1}^{m} v_{\ell j}^2(\mathcal{W})\lambda_\ell(\mathcal{W}))$ and $\bar{\zeta}_{kj} = \sigma_k^2/(\sum_{\ell=1}^{m} v_{\ell j}^2(\mathcal{W})\sigma_\ell^2)$. Under the assumptions of Theorem 1, as $d \to \infty$, for $k, j = 1 \ldots, m$,*

(i) $r(\hat{w}_k, w_j) \to v_{kj}(\mathcal{W})\zeta_{kj}^{1/2}$ in probability ;

(ii) $\lim_{d\to\infty} Corr(\hat{w}_{k*}, w_{j*} \mid W_1) = v_{kj}(\mathcal{W})\bar{\zeta}_{kj}^{1/2}$.

Remark 3. In the special case, $m = 1$, both the sample and prediction scores of the first principal component are perfectly correlated with the true scores, in the limit. Specifically, Theorem 2 implies that $|r(\hat{w}_1, w_1)| \to 1$ in probability and $|\text{Corr}(\hat{w}_{k*}, w_{j*})| \to 1$ as $d \to \infty$.

Remark 4. The somewhat complex limiting quantity $v_{kj}(\mathcal{W})\zeta_{kj}^{1/2}$ is an artifact of the fixed sample size. To simplify the expression for the case $k = j$, write

$$\left(v_{kk}(\mathcal{W})\zeta_{kk}^{1/2}\right)^2 = \frac{1}{1+\xi_k(\mathcal{W})}, \quad \xi_k(\mathcal{W}) = \sum_{\ell \neq k} v_{\ell k}^2(\mathcal{W})\frac{\lambda_\ell(\mathcal{W})}{\lambda_k(\mathcal{W})}.$$

Note that $\mathcal{W} = W_1 W_1^{\mathrm{T}}$ is proportional to the sample covariance matrix of the first m true scores, and that $v_{kk}(\mathcal{W})$ is the inner product between the kth sample and theoretical principal component directions of the data set W_1, where the number of variables, m, is smaller than the sample size n. Therefore, we expect that $|v_{kk}(\mathcal{W})| \approx 1$ and $\xi_k(\mathcal{W}) \approx 0$ for large sample size n. Taking the additional limit $n \to \infty$, the results in Theorem 2 become more interpretable:

$$|r(\hat{w}_k, w_j)| \to 1_{(k=j)} \text{ in probability, and } |\text{Corr}(\hat{w}_{k*}, w_{j*})| \to 1_{(k=j)},$$

as $d \to \infty, n \to \infty$ (limits are taken progressively).

Remark 5. What is the correlation coefficient $r(\hat{w}_k, w_k)$ for $k > m$ in the limit $d \to \infty$? In an attempt to answer this question, we note $\hat{w}_k = (n\hat{\lambda}_k)^{1/2}\hat{v}_k$, $\hat{v}_k = v_k(\mathcal{X}^T\mathcal{X})$ and $\mathcal{X}^T\mathcal{X} = \sum_{i=1}^d w_i w_i^T$. Thus,

$$r(\hat{w}_k, w_k) = w_k^T v_k(\sum_{i=1}^d w_i w_i^T)/\sqrt{\lambda_k},$$

and it is natural to guess that the dependence of $\hat{v}_k$ on any w_i, including the case $i = k$, would diminish as d tends to infinity. In fact, $d^{-1}\mathcal{X}^T\mathcal{X}$ converges to the rank-m matrix $S_0 := W_1^T W_1 + \tau^2 I_n$ (Jung et al., 2012), and w_k and S_0 are independent. Thus, it is reasonable to conjecture that $\lim_{d\to\infty} \mathrm{E}[r(\hat{w}_k, w_k)] = 0$, for $k > m$. Unfortunately, in the limit $d \to \infty$, the kth, $k > m$, eigenvector of $d^{-1}\mathcal{X}^T\mathcal{X}$ becomes an arbitrary choice in the left null space of W_1. Due to this non-unique eigenvector, the inner product $w_k^T v_k(S_0)$ is not defined, and consequently discussing the convergence of $r(\hat{w}_k, w_k)$ is somewhat demanding. We numerically confirm the conjecture in Section 4.1.

2.4   Inconsistency of the direction and variance estimators

The findings in the previous subsection may be summarized as that the first m principal component scores convey about the same visual information as

the true values when displayed. (The information is further honed by the bias adjustment in Section 3.) In a practical point of view, the scores and their graph matter the most.

On the other hand, a quite different conclusion about the standard principal component analysis is made when the estimator $\hat{u}_i$ is of interest. The asymptotic behavior of the direction $\hat{u}_i$ as well as the variance estimator $\hat{\lambda}_i$ are obtained as a special case of Lemma 1. Under our model,

$$
(\hat{u}_i^{\mathrm{T}} u_i, d^{-1} n \hat{\lambda}_i) \to
\begin{cases}
(\rho_i^{-1} v_{ii}(\mathcal{W}), \lambda_i(\mathcal{W}) + \tau^2), & i = 1, \ldots, m; \\[2mm]
(0, \tau^2), & i = m+1, \ldots, n.
\end{cases}
\tag{2.12}
$$

in probability as $d \to \infty$ (n is fixed).

The variance estimator $\hat{\lambda}_i$, for $i \leqslant m$, is asymptotically proportionally-biased. Specifically, $\hat{\lambda}_i / \lambda_i \to (\lambda_i(\mathcal{W}) + \tau^2)/(n\sigma_i^2)$ in probability as $d \to \infty$. Thus by using a classical result on the expansion of the eigenvalues of $\mathcal{W}$ for large n,

$$
E(\hat{\lambda}_i / \lambda_i) \to 1 + \frac{1}{n} \left[\sum_{j \neq i}^{m} \frac{\sigma_j^2}{\sigma_i^2 - \sigma_j^2} + \frac{\tau^2}{\sigma_i^2} \right] + O(n^{-2}),
$$

as $d \to \infty$. Note that even when $m = 1$, the bias is still of order n^{-1}. This proportional bias may be empirically adjusted, using good estimates of σ_i^2 and τ^2. We do not pursue it here. Note that all empirical principal component variances, for $i > m$, converge to τ^2/n, when scaled by d, and thus do not reflect any information of the population.

The result (2.12) also shows that the direction estimator $\hat{u}_i$ is inconsistent and asymptotically-biased, compared to u_i. The estimator $\hat{u}_i$ is closer to u_i when $\rho_i^{-1}|v_{ii}(\mathcal{W})|$ is closer to 1. It is impossible to achieve $\rho_i^{-1}|v_{ii}(\mathcal{W})| \to 1$ since for finite n, both $|v_{ii}(\mathcal{W})|$ and ρ_i^{-1} are strictly less than 1. Although the "angle" between $\hat{u}_i$ and u_i is quantified in (2.12), the theorem itself is useless in adjusting the bias. This is because that the direction to which $\hat{u}_i$ moves away from u_i is random, i.e. uniformly distributed; see Wang & Fan (2017) for the limiting distribution of $\hat{u}_i$ under a general asymptotic scenario of $d/n \to \infty$, while $d/(n\lambda_i)^{-1}$ is bounded.

In short, while the bias in the principal component direction is challenging to remove, the bias in the sample and prediction scores can be quantified and removed.

3. Bias-adjusted scores

In this section, we describe and compare several choices for the estimation of the *bias-adjustment factor* ρ_i. Note that both sample and prediction scores are rotated by the same direction and amount, specified in the matrix R. For applications requiring score matching (e.g., classification rules trained on the sample scores or the ancestry estimation discussed in the introduction), coordinate-free methods are often used and there is less practical

advantage in estimating R. We focus on adjusting the scores by estimating ρ_i.

Suppose that the number of effective principal components, m, is pre-specified or estimated in advance. Our first estimator is obtained by replacing τ^2 and $\lambda_i(\mathcal{W})$ in $\rho_i = \sqrt{1 + \tau^2/\lambda_i(\mathcal{W})}$ with reasonable estimators. In particular, we set

$$\tilde{\tau}^2 = \frac{\sum_{i=m+1}^n \hat{\lambda}_i}{n-m} \frac{n}{d}, \quad \tilde{\lambda}_i(\mathcal{W}) = d^{-1}n\hat{\lambda}_i - \tau^2, \tag{3.13}$$

and

$$\tilde{\rho}_i = \sqrt{1 + \tilde{\tau}^2/\tilde{\lambda}_i(\mathcal{W})}, \quad (i = 1, \ldots, m). \tag{3.14}$$

This simple estimator $\tilde{\rho}_i$ is in fact consistent.

Corollary 1. *Suppose the assumptions of Lemma 1 are satisfied. Let $d \to \infty$. For $i = 1, \ldots, m$, conditional to W_1, $\tilde{\tau}^2$, $\tilde{\lambda}_i(\mathcal{W})$ and $\tilde{\rho}_i$ are consistent estimators of τ^2, $\lambda_i(\mathcal{W})$ and ρ_i, respectively.*

Using (3.14), the bias-adjusted sample and prediction scores are $\hat{w}_i^{(\text{adj})} = \tilde{\rho}_i^{-1}\hat{w}_i$ and $\hat{w}_{i*}^{(\text{adj})} = \tilde{\rho}_i\hat{w}_{i*}$ for $i = 1, \ldots, m$. The sample and prediction scores matrices in (2.5) and (2.6) are then adjusted to, using $\tilde{S} = \text{diag}(\tilde{\rho}_\cdot, \ldots, \tilde{\rho}_m)$,

$$\widehat{W}_1^{(\text{adj})} = \tilde{S}^{-1}\widehat{W}_1, \quad \widehat{W}_*^{(\text{adj})} = \tilde{S}\widehat{W}_*. \tag{3.15}$$
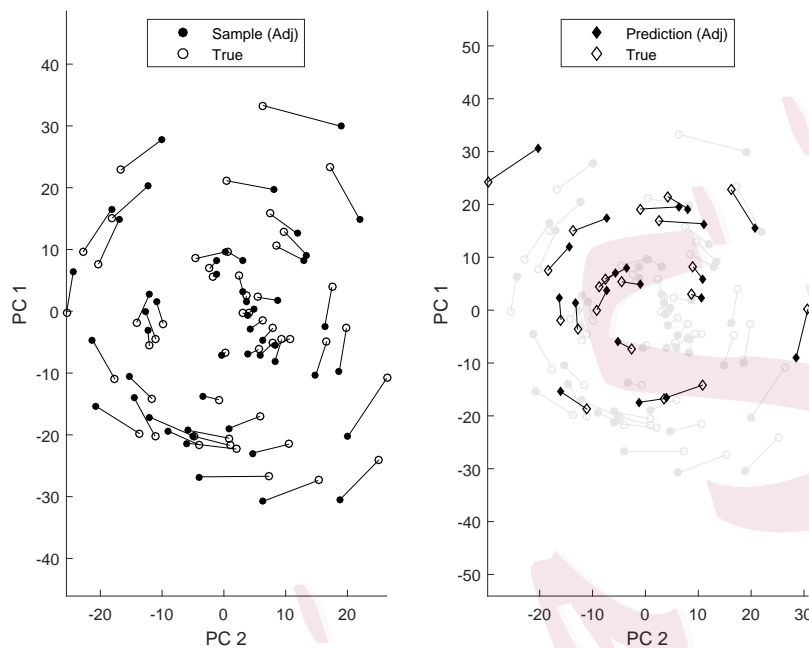
Figure 3: Bias-adjusted sample and prediction scores using (3.15) for the toy data introduced in Fig. 1. The estimates (3.14) are $(\tilde{\rho}_1, \tilde{\rho}_2) = (1.385, 1.546)$, very close to the theoretical values $(\rho_1, \rho_2) = (1.385, 1.557)$. Both sample and prediction scores are simultaneously rotated about 16 degrees clockwise.

An application of the above bias-adjustment procedure is exemplified in Fig. 3. There, the magnitudes of the sample and prediction scores are well-adjusted.

Our next proposed estimators are motivated by the well-known jack-

knife bias adjustment procedures and also by the leave-one-out cross-validation. For simplicity, assume $m = 1$. The bias-adjustment factor we aim to estimate is $\rho_1 = (1 + \tau^2/\|\xi_1\|_2^2)^{1/2}$, where $\xi_1 = d^{-1/2}w_1 = \sigma_1(z_{11}, \ldots, z_{1n})^{\mathrm{T}}$ is the scaled true scores for the first principal component.

Write, for each $j = 1, \ldots, n$, the jth scaled sample score as $\hat{\varpi}_{1j} = d^{-1/2}\hat{u}_1^{\mathrm{T}}X_j$, and the jth scaled prediction score as

$$\hat{\varpi}_{1(j)} = d^{-1/2}\hat{u}_{1(-j)}^{\mathrm{T}}X_j,$$

where $\hat{u}_{1(-j)}$ is the first principal component direction, computed from $\mathcal{X}_{(-j)}$, i.e., the data except the jth observation.

From Theorem 1, ρ_1 is the asymptotic bias-adjustment factor for $\hat{\varpi}_1$; $\hat{\varpi}_{1j} = \rho_1\varpi_{1j} + O_p(d^{-1/4})$. For $\hat{\varpi}_{1(j)}$, again applying Theorem 1, we get $\hat{\varpi}_{1(j)} = \rho_{1(-j)}^{-1}\varpi_{1j} + O_p(d^{-1/2})$, where $\rho_{1(-j)} = (1 + \tau^2/\|\varpi_{1(-j)}\|_2^2)^{1/2}$ is the bias-adjustment factor computed from $\mathcal{X}_{(-j)}$, using $\varpi_{1(-j)} = \sigma_1(z_{11}, \ldots, z_{1,j-1}, z_{1,j+1}, \ldots z_{1n})^{\mathrm{T}}$. To simplify terms, Taylor expansion is used to expand $\rho_{1(-j)}$ as a function of ϖ_{1j}^2/n, resulting in

$$\rho_{1(-j)} = \left(1 + \frac{\tau^2/n}{\|\varpi_1\|_2^2/n - \varpi_{1j}^2/n}\right)^{1/2} = \rho_1 + \frac{1}{2\rho_1}\frac{\|\varpi_1\|_2^2/n}{\tau^2}\frac{\varpi_{1j}^2}{n} + O_p(\frac{1}{n^2}).$$

$$(3.16)$$

Using the approximation

$$\rho_1\rho_{1(-j)} \approx \rho_1^2 + \frac{\|\varpi_1\|_2^2}{2\tau^2}\frac{\varpi_{1j}^2}{n^2}$$

given by (3.16), we write the ratio of the sample and prediction scores to cancel out the unknown true score ϖ_{1j} as follows:

$$\left(\frac{\hat{w}_{1j}}{\hat{w}_{1(j)}}\right)^{1/2} = \left(\frac{\hat{\varpi}_{1j}}{\hat{\varpi}_{1(j)}}\right)^{1/2} \approx \rho_1.$$

Based on the above heuristic, we define the following estimators of the bias-adjustment factors:

$$\hat{\rho}_i^{(1)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\hat{w}_{ij}}{\hat{w}_{i(j)}}\right)^{1/2}, \tag{3.17}$$

$$\hat{\rho}_i^{(2)} = \left(\frac{\sum_{j=1}^{n} \hat{w}_{ij}}{\sum_{j=1}^{n} \hat{w}_{i(j)}}\right)^{1/2}, \tag{3.18}$$

$$\hat{\rho}_i^{(3)} = \left(\frac{\sum_{j=1}^{n} \hat{w}_{ij}^2}{\sum_{j=1}^{n} \hat{w}_{i(j)}^2}\right)^{1/4}. \tag{3.19}$$

In implementing the above estimators, we used absolute values of the sample and predicted scores. The estimator (3.19) is a ratio of the sample and prediction score variances, obtained by a leave-one-out estimation of prediction scores.

The estimators $\hat{\rho}_i^{(1)}$, $\hat{\rho}_i^{(2)}$, and $\hat{\rho}_i^{(3)}$ tend to overestimate ρ for small sample size n, as expected from (3.16). In our numerical experiments, these three estimators perform similarly.

4. Numerical studies

4.1 Simulations to confirm the asymptotic bias and near-perfect correlations

In this section, we compare the theoretical asymptotic quantities derived in Section 2.3 with their finite-dimensional empirical counterparts.

First, the theoretical values of the scaling bias ρ_i and the rotation matrix R in Theorem 1 are compared with their empirical counterparts. The empirical counterparts of the two matrices R, S are defined as the minimizer of the Procrustes problem

$$\min \left\| W_1 - \widehat{W}_1^{\mathrm{T}} S_0^{-1} R_0 \right\|_F^2, \tag{4.20}$$

with the constraint that S_0 is a diagonal matrix with positive entries and R_0 is an orthogonal matrix. The solutions are denoted by $\check{S} = \mathrm{diag}(\check{\rho}_1(W_1), \ldots, \check{\rho}_m(W_1))$ and $\check{R}$. For simplicity, we consider the $m = 2$ case, and parameterize R by the rotation angle, $\theta_R = \cos^{-1}(R_{1,1})$, and $\check{R}$ by $\check{\theta}_R = \cos^{-1}(\check{R}_{1,1})$. We compare θ_R with $\check{\theta}_R$ and $\rho_i(W_1)$ with $\check{\rho}_i(W_1)$, from a 2-component model with $(n, d) = (50, 5000)$ (precisely, the spike model with $m = 2$ and $\beta = 0.3$ in Section 4.2). Note that both the theoretical values and the best-fitted values depend on the true scores W_1. To capture the natural variation given by W_1, the experiment is repeated for 100 times. The results, summarized

in the top row of Fig. 4, confirm that the asymptotic statements in Theorem 1 approximately hold for finite dimensions. In particular, the rotation matrices R and $\check{R}$ are very close to each other. The Procrustes-fitted, or "best", $\check{\rho}_i$ tends to be larger than the asymptotic, or theoretical, ρ_i, especially for $i = 2$ (shown as $\bigcirc$ in Fig. 4) and for larger values of ρ_2. This is not unexpected. Larger values of ρ_2 are from smaller $\lambda_2(\mathcal{W})$. Take an extreme case where $\lambda_2(\mathcal{W}) = 0$, then by (2.7) in Theorem 1, the sample scores are of magnitude $d^{1/2}$ compared to the true scores. Thus, as $\lambda_2(\mathcal{W})$ decreases to 0, the Procrustes scaler $\check{\rho}_2$ empirically interpolates the finite-scaling case (2.5) to the diverging case (2.7) of Theorem 1.

Second, we compare the limit of correlation coefficients in Theorem 2 with finite-dimensional empirical correlations, $r(\hat{w}_k, w_k)$, for $k = 1, 2$. For the correlation coefficient of the prediction scores, we use the sample correlation coefficient between $(\hat{w}_{k*}, w_{k*})$, as an estimate of $\mathrm{Corr}(\hat{w}_{k*}, w_{k*} \mid W_1)$. The simulated results are shown in the bottom row of Fig. 4. The empirical correlation coefficients tend to be smaller than the theoretical counterparts, but both are higher for stronger "signal strength" $n\sigma_k^2 = \mathrm{E}(\lambda_k(\mathcal{W}))$.

Third, from the same simulations, it can be checked that the kth, where $k > m$, sample scores are diverging, while the prediction scores are stable, as indicated in (2.7) and (2.8). To confirm this, we choose $k = 3$ and for each
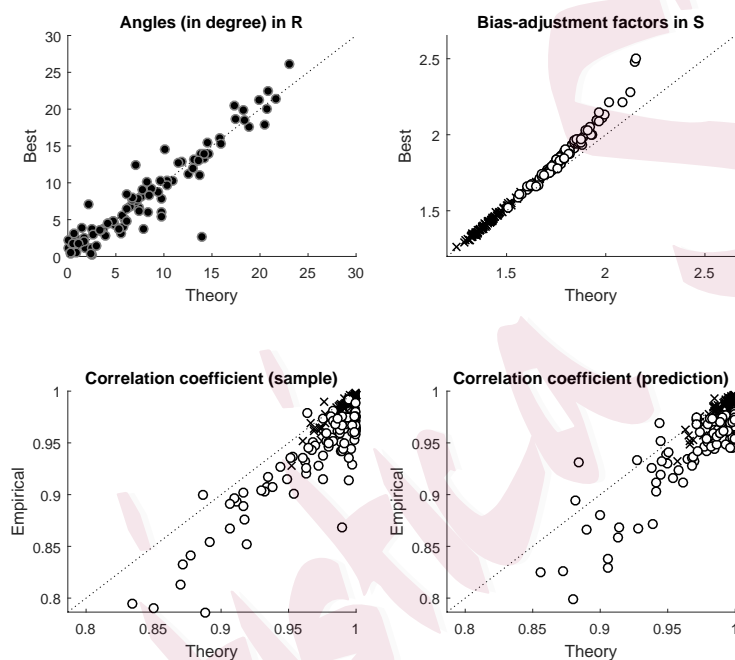
Figure 4: (Top row) Theoretical rotation angles θ_R and bias-adjustment factors ρ_1 ($\times$), ρ_2 ($\bigcirc$), compared with the best-fitting Procrustes counterparts $(\check{\theta}_R, \check{\rho}_i(W_1))$. (Bottom row) Empirical correlations compared with their limits in Theorem 2.

	Sample scores	Prediction scores
Variance	120.7(4.4)	1.38(0.2)
Corr. Coef.	-0.0024(0.2)	-0.004(0.15)

Table 1: The kth sample and prediction scores (unadjusted) for the case $k > m$. Shown are the mean (standard deviation) of the variances and correlation coefficients to true scores, from 100 repetitions. The true variance is $\lambda_3 = \mathrm{Var}(w_{3*}) \approx 6.5$.

experiment, compute $\widehat{\mathrm{Var}}(\hat{w}_3)$, the sample variance of the sample scores, and an approximation of $\mathrm{Var}(\hat{w}_{3*})$. The results are shown in Table 1. As expected, the sample scores are grossly inflated, while the prediction scores are stable. Finally, the conjecture in Remark 5 is also empirically checked; Table 1 also shows that for large d, the sample (or prediction) and true scores for the kth, $k > m$, component are nearly uncorrelated.

4.2   Numerical performance of the bias-adjustment factor estimation

We now test our estimators of the bias-adjustment factor ρ_i, using the following data-generating models with $m = 2$.

The first one is called a *spike model*. We sample from the d-dimensional

zero-mean normal distribution where the first two largest eigenvalues of the covariance matrix are $\lambda_i = \sigma_i^2 d$, for $i = 1, 2$, where $(\sigma_1^2, \sigma_2^2) = (0.02, 0.01)$. The rest of eigenvalues are slowly-decreasing. In particular, $\lambda_i = \tau i^{-\beta}$, where $\tau = [\sum_{i=3}^{d} i^{-\beta}/(d-2)]^{-1}$. We set $\beta = 0.3$ or 0.5. This spike model has more than two unique principal components for each fixed dimension, but in the limit $d \to \infty$, only the first two principal components are useful.

The second model is a *mixture model*. Let μ_g ($g = 1, 2, 3$) be d-dimensional vectors, the elements of which are randomly drawn from $\{-a, 0, a\}$ with replacement for a given $a > 0$, then assumed as fixed quantities. Given μ_g's we sample from the mixture model $X \mid G = g \sim N(\mu_g, \mathbb{I}_d)$, $P(G = g) = p_g > 0$, $\sum_{g=1}^{3} p_g = 1$. We set $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$. It can be checked that $\text{Cov}(X)$ satisfies the assumption of the 2-component model in (A1)–(A4).

For various cases of high-dimension, low-sample-size situations, ranging $d = 5,000$ to $20,000$ and $n = 50$ to 100, random samples from each of these models are generated. For each case, the theoretical quantity $\rho_i = \rho_i(W_1)$ and the best-fitted Procrustes scaler $\check{\rho}_i = \check{\rho}_i(W_1)$ are computed. These quantities depend on the $m \times n$ random matrix W_1. The mean and standard deviation of ρ_i (from 100 repetitions) are shown in the first column of Table 2. As expected, the theoretical value ρ_i depends on the sample

size n; large sample size decreases the bias, $\mathrm{E}(\rho_i)$, and also decreases the variance $\mathrm{Var}(\rho_i)$.

The mean of the best-fitted scaler $\check{\rho}_i$ $(i = 1)$ is displayed in the second column of the table. While they are quite close to the theoretical counterpart, $\check{\rho}_i$s are significantly larger for the mixture model, whose signal-to-noise ratio is smaller than the spike model, and for the not-so-large dimension $d = 5,000$. This is not unexpected, since the theoretical values are also based on the dimension-increasing asymptotic arguments.

We further compute the proposed estimators of ρ_i, given in (3.14), (3.17)–(3.19). We also compute the estimator derived from Lee et al. (2010), which is the square-root of the reciprocal of the shrinkage factor, obtained by numerical iterations, denoted by $\hat{d}_\nu$ in Lee et al. (2010). (The relation of Lee et al. (2010) to our work is further discussed in Section 5.) All of the methods considered provide accurate estimates of the theoretical quantity ρ_i. We omit the numerical results from the estimators (3.18) and (3.19), as their performances are very close to those from (3.17). The supplementary material contains an extended table of Table 2, including the case for ρ_2.

			ρ_1				
	d	n	Theory	Best	Asymp.	Jackknife	LZW
	5000	50	1.41 (0.07)	1.42	1.40	1.43	1.41
Spike model	10000	50	1.42 (0.06)	1.43	1.42	1.44	1.42
$\beta = 0.3$	10000	100	1.23 (0.03)	1.23	1.23	1.24	1.23
	20000	100	1.23 (0.02)	1.23	1.23	1.24	1.23
	5000	50	1.42 (0.08)	1.45	1.41	1.45	1.40
Spike model	10000	50	1.43 (0.07)	1.45	1.43	1.46	1.42
$\beta = 0.5$	10000	100	1.22 (0.02)	1.23	1.22	1.23	1.21
	20000	100	1.23 (0.02)	1.23	1.23	1.24	1.22
	5000	50	2.06 (0.06)	2.22	1.92	2.14	2.00
Mixture model	10000	50	2.09 (0.06)	2.17	1.98	2.14	2.02
$a = 0.15$	10000	100	1.63 (0.02)	1.67	1.61	1.65	1.63
	20000	100	1.64 (0.02)	1.66	1.62	1.66	1.63

Table 2: Simulation results from 100 repetitions. "Theory" is mean (standard deviation) of ρ_i; "Best" is $\check{\rho}_i$ (4.20); "Asymp." is $\tilde{\rho}_i$ (3.14); "Jackknife" is $\hat{\rho}_i^{(1)}$ (3.17); "LZW" is from Lee et al. (2010). Averages are shown for the latter four columns. The standard errors of the quantities in estimation of ρ_i are at most 0.04.

4.3   Bias-adjustment improves classification

Our last simulation study is an application of the bias-adjustment procedure to classification. Our training and testing data, each with sample size 100, are sampled from the mixture model with three groups, as described in Section 4.2. As frequently used in practice (Adam et al., 2008), dimension reduction by the standard principal component analysis is performed first, then a classification rule by the support vector machine (SVM, Cristianini & Shawe-Taylor, 2000) is trained on the sample principal component scores. In this simulation, we fix $m = 2$ and $d = 5000$. We compare the training and testing missclassification error rates (estimated by 100 repetitions) of the SVMs trained (and tested) either on the unadjusted sample and prediction scores, $\widehat{W}_1$ and $\widehat{W}_\star$, or on the bias-adjusted sample and prediction scores, $\widehat{W}_1^{(\mathrm{adj})}$ and $\widehat{W}_\star^{(\mathrm{adj})}$ in (3.15). The estimated error rates are shown in Table 3. It is clear that the use of bias-adjusted scores greatly improves the performance of classification.

To better understand the huge improvement of classification performances, we plot the sample and prediction scores that are inputs of the classifier. In Fig. 5, the classifier is estimated from the the sample scores (symbol $\bigcirc$) and is used to classify future observations, i.e. the prediction scores (symbol $\times$). Due to the scaling bias, the unadjusted sample and

	Unadjusted scores	Bias-adjusted scores
Training Error	0.04(0.02)	0.07(0.03)
Testing Error	21.4(1.33)	1.98(0.23)

Table 3: Means (standard errors) of Missclassification error rates (in percent).

prediction scores are of different scales (shown in the left panel), and classification is bound to fail. On the other hand, the proposed bias-adjustment, shown in the right panel, works well for this data set, leading to a better classification performance.

5. Discussion

The standard principal component analysis is shown to be useful in the dimension reduction of data from the m-component models with diverging variances. In particular, in the high-dimension, low-sample-size asymptotic scenario we reveal that the sample and prediction scores have systematic biases that can be consistently adjusted. We propose several estimators of the scaling bias, while there is no compelling reason to adjust rotational bias. The amount of bias is large when the sample size is small and when the variance of accumulated noise is large compared to the variances of the
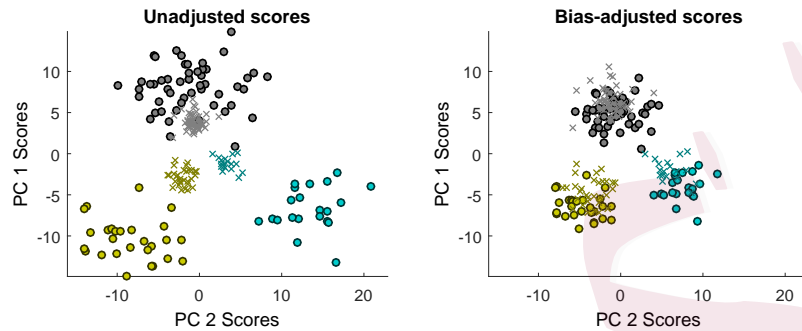
Figure 5: Bias-adjusted scores from the mixture models greatly improve the classification performance. Different colors correspond to different groups. Symbol ◯ represents the sample scores (unadjusted in the left, adjusted in the right); symbol × represents the prediction scores.

first m components.

Lee et al. (2010) discussed adjusting bias in the prediction of principal components, based on the random matrix theory and the asymptotic scenario of $d/n \to \gamma \in (0, \infty)$, $n \to \infty$. They showed that the prediction scores tend to be smaller than the sample scores, and the ratio of the shrinkage is asymptotically $\mathrm{sd}(\hat{w}_{i1})/\mathrm{sd}(\hat{w}_{i*}) \approx \rho_i^{(\mathrm{LZW})} = \frac{\lambda_i - 1}{\lambda_i + \gamma - 1}$. This "shrinkage factor" $\rho_i^{(\mathrm{LZW})}$ corresponds to the squared reciprocal of our scaling bias, ρ_i^{-2}. Our work can be thought of as an extension of Lee et al. (2010) from the

asymptotic regime $d \asymp n$ to the high-dimension, low-sample-size situations (see also Lee et al. (2014); Dey & Lee (2019)). Finally, we note that in the asymptotic scenario of Lee et al. (2010, 2014) and Dey & Lee (2019) there is no rotational bias. This is because in their limit the sample size is infinite. We show that the rotational bias is universal to both sample and prediction scores and is of order $n^{-1/2}$.

Supplementary Materials

The online supplementary materials contain proofs of all results and a table summarizing simulation results.

Acknowledgements

References

ABRAHAM, G. & INOUYE, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PloS one* **9**, e93766.

ADAM, C. D., SHERRATT, S. L. & ZHOLOBENKO, V. L. (2008). Classification and individualisation of black ballpoint pen inks using principal component analysis of UV–vis absorption

REFERENCES

spectra. *Forensic Sci. Int.* **174**, 16–25.

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math.
Stat.* **34**, 122–148.

AOSHIMA, M., SHEN, D., SHEN, H., YATA, K., ZHOU, Y.-H. & MARRON, J. (2018). A survey
of high dimension low sample size asymptotics. *Aust. N. Z. J. Stat* **60**, 4–19.

CRISTIANINI, N. & SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines.*
Cambridge University Press.

DEY, R. & LEE, S. (2019). Asymptotic properties of principal component analysis and
shrinkage-bias adjustment under the generalized spiked population model. *J. Multivar.
Anal* **173**, 145–164.

FAN, J., HAN, F. & LIU, H. (2014). Challenges of big data analysis. *Natl. Sci. Rev.* **1**, 293–314.

FAN, J., LIAO, Y. & MINCHEVA, M. (2013). Large covariance estimation by thresholding
principal orthogonal complements. *J. R. Stat. Soc. B* **75**, 603–680.

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space.
J. R. Stat. Soc. B **70**, 849–911.

HELLTON, K. H. & THORESEN, M. (2017). When and why are principal component scores a
good tool for visualizing high-dimensional data? *Scand. J. Stat.* **44**, 581–597.

JACKSON, J. E. (2005). *A user's guide to principal components*, vol. 587. John Wiley & Sons.

JOHNSTONE, I. M. & LU, A. Y. (2009). On Consistency and Sparsity for Principal Components

REFERENCES

Analysis in High Dimensions. *J. Am. Stat. Assoc.* **104**, 682–693.

JUNG, S., AHN, J. & LEE, M. H. (2018). On the number of principal components in high dimensions. *Biometrika* **105**, 389–402.

JUNG, S. & MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Stat.* **37**, 4104–4130.

JUNG, S., SEN, A. & MARRON, J. (2012). Boundary behavior in High Dimension, Low Sample Size asymptotics of PCA. *J. Multivar. Anal.* **109**, 190–203.

LEE, S., ZOU, F. & WRIGHT, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Stat.* **38**, 3605.

LEE, S., ZOU, F. & WRIGHT, F. A. (2014). Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika* **101**, 484.

LI, Q., CHENG, G., FAN, J. & WANG, Y. (2017). Embracing the blessing of dimensionality in factor models. *J. Am. Stat. Assoc.* **113**, 380–389.

LI, Q., SHANG, L., GAO, T., ZHANG, L., OU, T., HUANG, G., CHEN, C. & LI, C. (2014). Use of principal component scores in multiple linear regression models for simulation of chlorophyll-a and phytoplankton abundance at a karst deep reservoir, southwest of China. *Acta Ecologica Sinica* **34**, 72–78.

MARCUS, J. H., POSTH, C., RINGBAUER, H., LAI, L., SKEATES, R., SIDORE, C., BECKETT, J., FURTWÄNGLER, A., OLIVIERI, A., CHIANG, C. W. et al. (2020). Genetic history from

REFERENCES

the Middle Neolithic to present on the Mediterranean island of Sardinia. *Nat. Commun.*
11, 1–14.

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance
model. *Stat. Sin.* **17**, 1617–1642.

SHEN, D., SHEN, H., ZHU, H. & MARRON, J. (2016). The statistics and mathematics of high
dimension low sample size asymptotics. *Stat. Sin.* **26**, 1747.

SUNDBERG, R. & FELDMANN, U. (2016). Exploratory factor analysis-parameter estimation and
scores prediction with high-dimensional data. *J. Multivar. Anal.* **148**, 49–59.

WANG, C., ZHAN, X., LIANG, L., ABECASIS, G. R. & LIN, X. (2015). Improved ancestry
estimation for both genotyping and sequencing data using projection procrustes analysis
and genotype imputation. *Am. J. Hum. Genet.* **96**, 926–937.

WANG, W. & FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional
spiked covariance. *Ann. Stat.* **45**, 1342.

ZHAN, X., LARSON, D. E., WANG, C., KOBOLDT, D. C., SERGEEV, Y. V., FULTON, R. S.,
FULTON, L. L., FRONICK, C. C., BRANHAM, K. E., BRAGG-GRESHAM, J. et al. (2013).
Identification of a rare coding variant in complement 3 associated with age-related macular
degeneration. *Nat. Genet.* **45**, 1375–1379.

ZHANG, D., DEY, R. & LEE, S. (2020). Fast and robust ancestry prediction using principal
component analysis. *Bioinformatics* **36**, 3439–3446.

<div style="text-align: right;">

REFERENCES

</div>

ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2006). Sparse Principal Component Analysis. *J. Comp. Graph. Stat.* **15**, 265–286.

Seoul National University

E-mail: sungkyu@snu.ac.kr