Statistica Sinica **35** (2025), 1069-1090
doi:https://doi.org/10.5705/ss.202022.0319

AUTOMATIC SPARSE PCA FOR HIGH-DIMENSIONAL DATA

Kazuyoshi Yata* and Makoto Aoshima

University of Tsukuba

Abstract: Sparse principal component analysis (SPCA) methods have proven to efficiently analyze high-dimensional data. Among them, threshold-based SPCA (TSPCA) is computationally more cost-effective than regularized SPCA, based on L1 penalties. We herein present an investigation of the efficacy of TSPCA for high-dimensional data settings and illustrate that, for a suitable threshold value, TSPCA achieves satisfactory performance for high-dimensional data. Thus, the performance of the TSPCA depends heavily on the selected threshold value. To this end, we propose a novel thresholding estimator to obtain the principal component (PC) directions using a customized noise-reduction methodology. The proposed technique is consistent under mild conditions, unaffected by threshold values, and therefore yields more accurate results quickly at a lower computational cost. Furthermore, we explore the shrinkage PC directions and their application in clustering high-dimensional data. Finally, we evaluate the performance of the estimated shrinkage PC directions in actual data analyses.

Key words and phrases: Clustering, large p small n, PCA consistency, shrinkage PC directions, thresholding.

1. Introduction

High-dimensional, low-sample-size (HDLSS) data scenarios exist in many areas of modern science including genomics, medical imaging, text recognition, and finance. In recent years, substantial work has been conducted on HDLSS asymptotic theory, wherein the sample size n is fixed or $n/d \to 0$ is used as the data dimension $d \to \infty$. For principal component analysis (PCA), Jung and Marron (2009) and Yata and Aoshima (2009) investigated inconsistency properties for both the eigenvalues and principal component (PC) directions in a sample covariance matrix. Yata and Aoshima (2012) developed a new PCA method called the *noise-reduction methodology* and reported consistent estimators for both eigenvalues and PC directions in addition with the PC scores using this method. Sparse PCA (SPCA) methods have been investigated in several studies. For example, Zou and Hastie (2006), Shen and Huang (2008), and Lee, Huang and Hu (2010) considered a regularized SPCA (RSPCA) based on L1 penalties under high-dimensional settings. Johnstone and Lu (2009) proposed a

*Corresponding author. E-mail: yata@math.tsukuba.ac.jp

thresholded SPCA (TSPCA) and presented a consistency property of the TSPCA when $n/d \to 0$. Further, Shen, Shen and Marron (2013) showed that the PC directions obtained by RSPCA and TSPCA are consistent when $d \to \infty$ while n is fixed. In addition, Paul and Johnstone (2007) developed an augmented SPCA method, and Ma (2013) proposed an iterative thresholding procedure for PC directions. In this study, we focused on TSPCA under high-dimensional settings.

Suppose that we have a $d \times n$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, where $\boldsymbol{x}_i = (x_{i(1)}, \ldots, x_{i(d)})^T$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) as a d-dimensional distribution with mean $\boldsymbol{\mu}$ and (non-negative definite) covariance matrix $\boldsymbol{\Sigma}$. We express the eigen-decomposition of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, where $\boldsymbol{\Lambda}$ represents a diagonal matrix of the eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_d (\geq 0)$, and $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d)$ is an orthogonal matrix of the corresponding eigenvectors. The sample covariance matrix is given by $\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T = (n-1)^{-1}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$, where $\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$ and $\overline{\boldsymbol{X}} = \bar{\boldsymbol{x}}\boldsymbol{1}_n^T$ with $\boldsymbol{1}_n = (1, \ldots, 1)^T$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d \geq 0$ be the eigenvalues of $\boldsymbol{S}$, and let $\hat{\boldsymbol{h}}_j$, $j = 1, \ldots, d$ be the corresponding eigenvectors $\hat{\boldsymbol{h}}_j^T \hat{\boldsymbol{h}}_{j'} = \delta_{jj'}$, where $\delta_{jj'}$ is the Kronecker delta. Thus, the eigen-decomposition of $\boldsymbol{S}$ is $\boldsymbol{S} = \sum_{s=1}^d \hat{\lambda}_s \hat{\boldsymbol{h}}_s \hat{\boldsymbol{h}}_s^T$. We assume that $\boldsymbol{h}_j^T \hat{\boldsymbol{h}}_j \geq 0$ for all j without the loss of generality. We now consider the $n \times n$ dual-sample covariance matrix defined by $\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$. Here, $\boldsymbol{S}$ and $\boldsymbol{S}_D$ share nonzero eigenvalues. Let the eigen-decomposition of $\boldsymbol{S}_D$ be $\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$, where $\hat{\boldsymbol{u}}_j$ denotes an eigenvector corresponding to $\hat{\lambda}_j$ and $\hat{\boldsymbol{u}}_j^T \hat{\boldsymbol{u}}_{j'} = \delta_{jj'}$. Furthermore, $\hat{\boldsymbol{h}}_j$ can be calculated as $\hat{\boldsymbol{h}}_j = \{(n-1)\hat{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j$.

Johnstone (2001), Baik and Silverstein (2006), and Paul (2007) considered a spiked model for the eigenvalues.

$$\begin{aligned} &\lambda_j\ (> \kappa), \quad j = 1, \ldots, m, \text{ are fixed (not depending on } d) \\ &\text{and } \lambda_{m+1} = \cdots = \lambda_d = \kappa. \end{aligned} \tag{1.1}$$

Here, m represents a fixed positive integer and $\kappa\ (> 0)$ represents a fixed constant. Under (1.1), the asymptotic behavior of the eigenvalues of $\boldsymbol{S}$ was studied when both d and n increased at the same rate, that is, from $n/d \to \gamma > 0$. Details under the Gaussian assumptions were reported by Johnstone (2001), Johnstone and Lu (2009), and Paul (2007). Further, Baik and Silverstein (2006) and Lee, Zou and Wright (2010) reported details under the non-Gaussian but i.i.d. assumptions as in (2.4). For review, the authors direct readers to Paul and Aue (2014). The condition $\lambda_{m+1} = \cdots = \lambda_d = \kappa$ is strict for the latter part of (1.1). Without assuming that $\lambda_{m+1} = \cdots = \lambda_d = \kappa$ in (1.1), Bai and Ding (2012) estimated the forward eigenvalues. However, the former part of (1.1) is a strict condition because the eigenvalues depend on d, and it is probable that $\lambda_j \to \infty$ for the first several js when $d \to \infty$. Details were provided by Fan, Liao and Mincheva

(2013), Jung and Marron (2009), Onatski (2012), Shen et al. (2016), Wang and Fan (2017), and Yata and Aoshima (2012, 2013). They considered spiked models such as

$$\lambda_j = \kappa_j d^{\alpha_j} \ (j = 1, \ldots, m) \quad \text{and} \quad \lambda_j = \kappa_j \ (j = m+1, \ldots, d). \tag{1.2}$$

Here, κ_j (> 0) and α_j $(\alpha_1 \geq \cdots \geq \alpha_m > 0)$ are fixed constants preserving the order that $\lambda_1 \geq \cdots \geq \lambda_d$. For example, Cai, Han and Pan (2020), Shen et al. (2016), Wang and Fan (2017), and Yata and Aoshima (2012) showed that

$$\frac{\hat{\lambda}_j}{\lambda_j} = 1 + \frac{\delta}{\lambda_j} + o_P(1) \ \text{ as } d \to \infty \text{ and } n \to \infty \text{ for } j = 1, \ldots, m$$

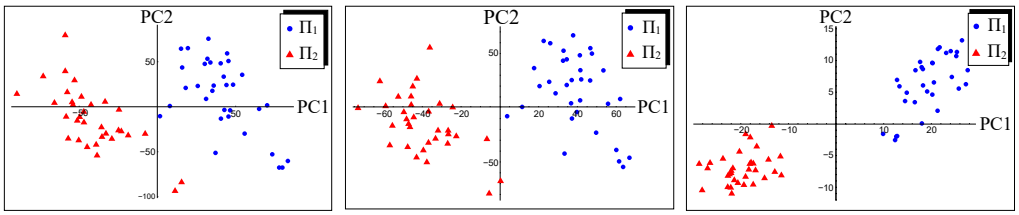
under (1.2), $d^{1-2\alpha_m}/n = o(1)$, and (2.4), where $\delta = \sum_{s=m+1}^{d} \lambda_s/(n-1)$. Further details are provided in Appendix E in the online supplementary material. Here, $\delta = O(d/n)$ for (1.2); if $\delta/\lambda_j \to \infty$, $\hat{\lambda}_j$ is strongly inconsistent in the sense that $\lambda_j/\hat{\lambda}_j = o_P(1)$. Jung and Marron (2009) and Yata and Aoshima (2009) have reported on the concept of strong inconsistency. Yata and Aoshima (2012) proposed a noise-reduction (NR) methodology that uses the geometric representation of high-dimensional eigenspaces to overcome the curse of dimensionality. If the NR method is used, λ_js can be estimated by

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{s=1}^{j} \hat{\lambda}_s}{n-j-1} \quad (j = 1, \ldots, n-1). \tag{1.3}$$

Yata and Aoshima (2012, 2013) showed the consistency as "$\tilde{\lambda}_j/\lambda_j = 1 + o_P(1)$" even when $\delta/\lambda_j \to \infty$. Section 2.3 provides further details. For PC direction $\hat{\boldsymbol{h}}_j$, under (1.2) and some regularity conditions, Yata and Aoshima (2012, 2013) and Shen et al. (2016) showed that $\mathrm{Angle}(\hat{\boldsymbol{h}}_j, \boldsymbol{h}_j) = \mathrm{Arccos}\{(1+\delta/\lambda_j)^{-1/2}\} + o_P(1)$ as $d \to \infty$ and $n \to \infty$ for $j = 1, \ldots, m$. If $\delta/\lambda_j \to \infty$, $\hat{\boldsymbol{h}}_j$ is strongly inconsistent in that $\mathrm{Angle}(\hat{\boldsymbol{h}}_j, \boldsymbol{h}_j) = \pi/2 + o_P(1)$. To overcome this inconvenience, Shen, Shen and Marron (2013) showed that estimators of $\boldsymbol{h}_1$ given by TSPCA and RSPCA have a consistency property and that they perform equivalently for high-dimensional data. However, TSPCA is easier to handle as compared to RSPCA. Appendix C in the online supplementary material provides further details in this regard. TSPCA can be summarized as follows: Let $\hat{\boldsymbol{h}}_j = (\hat{h}_{j(1)}, \ldots, \hat{h}_{j(d)})^T$ for all j. Given a sequence of threshold values $\zeta > 0$, we can define the thresholded entries as

$$\hat{h}_{j*(j')} = \begin{cases} \hat{h}_{j(j')} & \text{if } |\hat{h}_{j(j')}| \geq \min\{\zeta, \hat{h}_{j\max}\} \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } j' = 1, \ldots, d, \tag{1.4}$$

where $\hat{h}_{j\max} = \max_{s=1,\ldots,d} |\hat{h}_{j(s)}|$. Let $\hat{\boldsymbol{h}}_{j*} = (\hat{h}_{j*(1)}, \ldots, \hat{h}_{j*(d)})^T$. Then, the



(i) Conventional PCA (ii) TSPCA with $\zeta = 0.01$ (iii) TSPCA with $\zeta = 0.05$.

Figure 1. Scatter plots of the first two PC scores for (i) Conventional PCA, (ii) TSPCA with $\zeta = 0.01$, and (iii) TSPCA with $\zeta = 0.05$.

thresholded estimator of $\boldsymbol{h}_j$ is defined as

$$\hat{\boldsymbol{h}}_{j(\zeta)} = \frac{\hat{\boldsymbol{h}}_{j*}}{\|\hat{\boldsymbol{h}}_{j*}\|}, \tag{1.5}$$

where $\|\cdot\|$ denotes the Euclidean norm. Shen, Shen and Marron (2013) showed the consistency property

$$\text{Angle}(\hat{\boldsymbol{h}}_{1(\zeta)}, \boldsymbol{h}_1) = o_P(1) \text{ as } d \to \infty \text{ for some } \zeta$$

under (1.2), the Gaussian assumption, and some regularity conditions. Further, they showed consistency even when $\delta/\lambda_j \to \infty$. However, this estimator depends heavily on the choice of ζ.

We analyzed the microarray data given by Chiaretti et al. (2004) wherein the dataset comprises 12,625 ($= d$) genes. The dataset consists of two tumor cellular subtypes: B-cell (33 samples) and T-cell (33 samples). B-cell originally contained 95 samples. We used only the first 33 samples to maintain balance in the sample sizes with the T-cell. For the 66 ($= n$) samples, the scatter plots of the first two PC scores are shown in Figure 1 for three PCAs: (i) Conventional PCA, (ii) TSPCA with $\zeta = 0.01$, and (iii) TSPCA with $\zeta = 0.05$.

We have that $\pi/2 - \text{Angle}(\hat{\boldsymbol{h}}_{1(0.01)}, \hat{\boldsymbol{h}}_{2(0.01)}) = 0.154$, and $\pi/2 - \text{Angle}(\hat{\boldsymbol{h}}_{1(0.05)}, \hat{\boldsymbol{h}}_{2(0.05)}) = 0.266$. Further, we observed that the conventional PCA effectively classifies the dataset into two groups using the first PC score. The theoretical clarification was provided by Yata and Aoshima (2020). The TSPCA with $\zeta = 0.05$ separates them more clearly than that when using the conventional PCA. However, this may not hold consistency as "$\text{Angle}(\hat{\boldsymbol{h}}_{j(\zeta)}, \boldsymbol{h}_j) = o_P(1)$ for $j = 1, 2$" because $\pi/2 - \text{Angle}(\hat{\boldsymbol{h}}_{1(0.05)}, \hat{\boldsymbol{h}}_{2(0.05)}) = 0.266$. Thus, TSPCA provides preferable performance if one chooses a suitable ζ, while it may be inconsistent.

In this study, we investigated TSPCA in high-dimensional settings. The contributions of this study are as follows:

(I) We propose a new thresholding estimator for PC directions and present its consistency property that holds freely from threshold values.

(II) We propose a shrinkage PC direction and apply it to clustering.

The remainder of this paper is organized as follows. In Section 2, we present the asymptotic properties of the NR method under certain conditions. In Section 3, we modify the estimator of the PC directions derived using the NR method and propose a new TSPCA method. In Section 4, we investigate the performance of the proposed TSPCA method through simulations. In Section 5, we present the estimation of the shrinkage PC directions and their application to clustering. In Section 6, we investigate the performance of the estimated shrinkage PC directions in actual data analyses. In the online Supplementary Material, we apply the proposed TSPCA to the estimation of the intrinsic component of $\boldsymbol{\Sigma}$.

2. Preliminary

In this section, we lay out basic conditions, assumptions, and asymptotics for the construction of our PCA methods.

2.1. Strongly spiked eigenstructures

Aoshima and Yata (2018) provided two disjointed high-dimensional models: the strongly spiked eigenvalue (SSE) model defined as

$$\liminf_{d\to\infty} \frac{\lambda_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)} > 0 \tag{2.1}$$

and the non-SSE (NSSE) model defined by

$$\frac{\lambda_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)} \to 0 \quad \text{as } d \to \infty. \tag{2.2}$$

Notably, (2.2) is equivalent to "$\mathrm{tr}(\boldsymbol{\Sigma}^4)/\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 \to 0$ as $d \to \infty$". If $\alpha_1 \geq 0.5$ in (1.2), then the SSE model (2.1) holds. In contrast, if $\alpha_1 < 0.5$ in (1.2), the NSSE model (2.2) holds. We provide additional examples of the SSE model in Remark 6 in Section 5.2 and Appendix D in the online supplementary material. The two models are essential for statistical inference of high-dimensional data. We emphasize that it is not possible to ensure the accuracy of high-dimensional statistical inferences using the SSE model. The work by Aoshima and Yata (2018) can be referred to for further details. Aoshima and Yata (2018, 2019) proposed data-transformation methods based on the strongly spiked eigenstructures to overcome this inconvenience. Further, Yata and Aoshima (2020) considered clustering under the SSE model. The key to these references is the estimation of the strongly spiked eigenstructures. In this study, we focused on estimating the strongly spiked eigenstructures.

Let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$, where $\boldsymbol{\Sigma}_1 = \sum_{s=1}^{m} \lambda_s \boldsymbol{h}_s \boldsymbol{h}_s^T$ and $\boldsymbol{\Sigma}_2 = \sum_{s=m+1}^{d} \lambda_s \boldsymbol{h}_s \boldsymbol{h}_s^T$ with an unknown and positive fixed integer m (independent of d). Here, $\boldsymbol{\Sigma}_1$ is considered an intrinsic part and $\boldsymbol{\Sigma}_2$, a noise component. We assume the following model:

(C-i) $\lambda_1, \ldots, \lambda_m$ are distinct in that $\liminf_{d\to\infty}(\lambda_j/\lambda_{j'} - 1) > 0$ for $1 \le j < j' \le m$ when $m \ge 2$ and λ_m and λ_{m+1} satisfy

$$\liminf_{d\to\infty} \frac{\lambda_m^2}{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)} > 0 \quad \text{and} \quad \frac{\lambda_{m+1}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)} \to 0 \quad \text{as } d \to \infty.$$

Remark 1. (C-i) is an SSE model. The spiked model (1.2) with $\alpha_m \ge 0.5$ and $\kappa_j \neq \kappa_{j'}$ for $1 \le j \neq j' \le m$ satisfies (C-i). Aoshima and Yata (2018) provided a method to check whether the SSE model holds or not. Aoshima and Yata (2018) also provided a consistent estimator of m in (C-i).

Next, we consider a bounded condition for diagonal elements. Let $\sigma_{(j)} = \mathrm{Var}(x_{i(j)})$ for all j. Here, $\sigma_{(j)}$ represents the j-th diagonal element of $\boldsymbol{\Sigma}$. Let $\boldsymbol{A}_1 = \sum_{s=1}^{m} \boldsymbol{h}_s \boldsymbol{h}_s^T$ and $\boldsymbol{A}_2 = \boldsymbol{I}_d - \boldsymbol{A}_1 = \sum_{s=m+1}^{d} \boldsymbol{h}_s \boldsymbol{h}_s^T$, where $\boldsymbol{I}_d$ represents the d-dimensional identity matrix. Let $\boldsymbol{x}_{i,1} = (x_{i(1),1}, \ldots, x_{i(d),1})^T = \boldsymbol{A}_1 \boldsymbol{x}_i$ and $\boldsymbol{x}_{i,2} = (x_{i(1),2}, \ldots, x_{i(d),2})^T = \boldsymbol{A}_2 \boldsymbol{x}_i$ for all i, $\mathrm{Var}(\boldsymbol{x}_{i,s}) = \boldsymbol{\Sigma}_s$ for $s = 1, 2$. Let $\sigma_{(j),s} = \mathrm{Var}(x_{i(j),s})$, $s = 1, 2$, for all j. Notably, $\sigma_{(j),2} \le \sigma_{(j),1} + \sigma_{(j),2} = \sigma_{(j)}$ for all j. We assume the following bounded condition, as necessary:

(C-ii) $\liminf_{d\to\infty} \sigma_{(j),2} > 0$ and $\limsup_{d\to\infty} \sigma_{(j)} < \infty$ for all j.

The diagonal elements are typically bounded. Thus, (C-ii) generally holds. Under (C-ii), $\sigma_{(j),2} \in (0, \infty)$ as $d \to \infty$ for all j. Here, for function $f(\cdot)$, "$f(d) \in (0, \infty)$ as $d \to \infty$" implies that $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$. Then, $\mathrm{tr}(\boldsymbol{\Sigma}_2^2)/d \ge \sum_{s=1}^{d} \sigma_{(s),2}^2/d \in (0, \infty)$ and $\mathrm{tr}(\boldsymbol{\Sigma})/d \in (0, \infty)$ as $d \to \infty$ under (C-ii). Therefore, under (C-i) and (C-ii),

$$\limsup_{p\to\infty} \frac{\lambda_j}{d} < \infty \quad \text{and} \quad \liminf_{p\to\infty} \frac{\lambda_j}{d^{1/2}} > 0 \quad \text{for } j = 1, \ldots, m. \tag{2.3}$$

2.2. Assumptions of high-dimensional distributions

Let $\boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{1}_n^T = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}$. Then, $\boldsymbol{Z}$ represents a $d \times n$ sphered data matrix obtained from a distribution with an identity covariance matrix. Here, we write $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d)^T$ and $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{nj})^T$, $j = 1, \ldots, d$. $E(z_{ij} z_{ij'}) = 0$ ($j \neq j'$) and $\mathrm{Var}(\boldsymbol{z}_j) = \boldsymbol{I}_n$. For convenience, when $\lambda_j = 0$ for some j, we assume $\mathrm{Var}(\boldsymbol{z}_j) = \boldsymbol{I}_n$. Let $M_j = \mathrm{Var}(z_{ij}^2)$, $j = 1, \ldots, d$. If $\boldsymbol{X}$ is Gaussian, z_{ij}s are i.i.d. as the standard normal distribution $N(0, 1)$ and $M_j = 2$ for all j. We consider the following assumptions:

(A-i) $E(z_{ij_1}^2 z_{ij_2}^2) = 1$, $E(z_{ij_1} z_{ij_2} z_{ij_3}) = 0$, $E(z_{ij_1} z_{ij_2} z_{ij_3} z_{ij_4}) = 0$ for all $j_1 \neq j_2, j_3, j_4$; and $\limsup_{d\to\infty} M_j < \infty$ for all j.

Here, (A-i) naturally holds when $\boldsymbol{X}$ is Gaussian. Another example satisfying (A-i) is the case when $\boldsymbol{X}$ has a skew normal distribution (Aoshima and Yata, 2018, Rmk. S4.1). This assumption was provided by Bai and Saranadasa (1996), Chen and Qin (2010), and Aoshima and Yata (2011). In contrast, Baik and Silverstein (2006), Lee, Zou and Wright (2010), and Yata and Aoshima (2012) considered the following assumption:

$$z_{i1}, \ldots, z_{id} \text{ are independent (or i.i.d.).} \tag{2.4}$$

The first assumption in (A-i) is milder than (2.4).

Next, we consider a sub-exponential distribution. Let $\boldsymbol{h}_j = (h_{j(1)}, \ldots, h_{j(d)})^T$ for all j. From $x_{i(j),2} = \sum_{s=m+1}^{d} \lambda_s^{1/2} h_{s(j)} z_{is}$, we note that $\mathrm{Var}(x_{i(j),2} z_{ij'}) = \sigma_{(j),2}$ under (A-i) for $j = 1, \ldots, d$; $j' = 1, \ldots, m$. Under (C-ii), we consider the following assumption:

(A-ii) There exist positive (fixed) constants t_1 and t_2 such that

$$\limsup_{d\to\infty} E\{\exp(t x_{i(j),2}^2)\} < \infty \text{ for } |t| \le t_1 \text{ and all } j; \text{ and}$$
$$\limsup_{d\to\infty} E\{\exp(t x_{i(j),2} z_{ij'})\} < \infty \text{ for } |t| \le t_2, \text{ all } j \text{ and } j' = 1, \ldots, m.$$

If $\boldsymbol{X}$ is Gaussian, $x_{i(j),2}/\sigma_{(j),2}^{1/2}$ follows $N(0,1)$, and $x_{i(j),2}$ and $z_{ij'}$ are independent for all j and $j' = 1, \ldots, m$, so that $E\{\exp(t x_{i(j),2} z_{ij'})\} \le E\{\exp(|t| x_{i(j),2}^2 + |t| z_{ij'}^2)\} = E\{\exp(|\sigma_{(j),2} t| x_{i(j),2}^2/\sigma_{(j),2})\} E\{\exp(|t| z_{ij'}^2)\} = (1 - 2|\sigma_{(j),2} t|)^{-1/2} (1 - 2|t|)^{-1/2}$ if $|t| < \min\{1/2, 1/(2\sigma_{(j),2})\}$ for all j and $j' = 1, \ldots, m$. Thus, when $\boldsymbol{X}$ is Gaussian, (A-ii) holds under (C-ii). In Appendix B in the online supplementary material, we consider a milder assumption than (A-ii).

2.3. Estimation of eigenvalues and PC directions using the NR method

Eigenvalue estimation using the NR method is given by (1.3). The second term in (1.3) is the estimator of δ.

Proposition 1 (Aoshima and Yata, 2018; Yata and Aoshima, 2013). *Assume (A-i) and (C-i). Then, it holds for $j = 1, \ldots, m$ that $\tilde{\lambda}_j/\lambda_j = 1 + O_P(n^{-1/2})$ and*

$$\left(\frac{n}{M_j}\right)^{1/2} \left(\frac{\tilde{\lambda}_j}{\lambda_j} - 1\right) \Rightarrow N(0,1) \quad \textit{if} \quad \liminf_{d\to\infty} M_j > 0$$

as $d \to \infty$ and $n \to \infty$.

Notably, $\tilde{\lambda}_j$ is a consistent estimator of λ_j even when $\delta/\lambda_j \to \infty$. Thus, we recommend using $\tilde{\lambda}_j$ instead of $\hat{\lambda}_j$ for high-dimensional data. Wang and Fan (2017) applied $\tilde{\lambda}_j$s in "Shrinkage Principal Orthogonal complEment Threshold-

ing" when $d^{1/2} = o(\lambda_m)$. Anderson (1963) showed that $n^{1/2}(\hat{\lambda}_j/\lambda_j - 1) \Rightarrow N(0, 2)$ as $n \to \infty$ for Gaussian data when d is fixed. Thus, $\tilde{\lambda}_j$ has the same limiting distribution as in Proposition 1 for the Gaussian data.

When applying the NR method to the PC direction vector, we obtain

$$\tilde{\boldsymbol{h}}_j = \{(n-1)\tilde{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j \ \text{ for } j = 1, \ldots, n-1.$$

Notably, $\tilde{\boldsymbol{h}}_j = (\hat{\lambda}_j/\tilde{\lambda}_j)^{1/2}\hat{\boldsymbol{h}}_j$. From Aoshima and Yata (2018), we obtain the following result.

Proposition 2 (Aoshima and Yata, 2018). *Assume (A-i) and (C-i). Then, it holds for $j = 1, \ldots, m$ that $\tilde{\boldsymbol{h}}_j^T \boldsymbol{h}_j = 1 + O_P(n^{-1})$ as $d \to \infty$ and $n \to \infty$.*

Aoshima and Yata (2018, 2019) used $\tilde{\boldsymbol{h}}_j$ to develop a data transformation method under the SSE model in the two-sample test and high-dimensional classification. Here, $\tilde{\boldsymbol{h}}_j$ is not a unit vector and $\|\tilde{\boldsymbol{h}}_j\|^2 = \hat{\lambda}_j/\tilde{\lambda}_j$. Further, $\mathrm{Angle}(\tilde{\boldsymbol{h}}_j, \hat{\boldsymbol{h}}_j) = 0$, and therefore, $\mathrm{Angle}(\tilde{\boldsymbol{h}}_j, \boldsymbol{h}_j) = \mathrm{Angle}(\hat{\boldsymbol{h}}_j, \boldsymbol{h}_j)$. From Propositions 2, E.1 in Appendix E, and (G.1) in Appendix G in the online supplementary material, under (A-i) and (C-i), it holds for $j = 1, \ldots, m$ that

$$\|\tilde{\boldsymbol{h}}_j\|^2 = 1 + \frac{\delta}{\lambda_j}\{1 + O_P(n^{-1/2})\} + O_P(n^{-1}) \ \text{ and} \tag{2.5}$$
$$\|\tilde{\boldsymbol{h}}_j - \boldsymbol{h}_j\|^2 = \frac{\delta}{\lambda_j}\{1 + O_P(n^{-1/2})\} + O_P(n^{-1})$$

as $d \to \infty$ and $n \to \infty$. Considering Remark E.1 in Appendix E by noting that $2\{1 - (1 + \delta/\lambda_j)^{-1/2}\} < \delta/\lambda_j$, the norm loss of $\tilde{\boldsymbol{h}}_j$ is larger than that of $\hat{\boldsymbol{h}}_j$. However, from Proposition 2, $\tilde{\boldsymbol{h}}_j$ represents a consistent estimator of $\boldsymbol{h}_j$ considering the inner product even when $\delta/\lambda_j \to \infty$. Further, from Propositions 2 and (2.5), under (A-i) and (C-i), there exists a random d-dimensional vector $\tilde{\boldsymbol{v}}_j$ such that $\boldsymbol{h}_j^T\tilde{\boldsymbol{v}}_j = 0$,

$$\tilde{\boldsymbol{h}}_j = \{1 + O_P(n^{-1})\}\boldsymbol{h}_j + \tilde{\boldsymbol{v}}_j \text{ and } \|\tilde{\boldsymbol{v}}_j\|^2 = \frac{\delta}{\lambda_j}\{1 + O_P(n^{-1/2})\} + O_P(n^{-1}) \tag{2.6}$$

for $j = 1, \ldots, m$. The coefficient of $\boldsymbol{h}_j$ in $\tilde{\boldsymbol{h}}_j$ is asymptotically 1. In contrast, from (2.6), Proposition 1, and Preposition E.1, it holds that

$$\hat{\boldsymbol{h}}_j = \left(1 + \frac{\delta}{\lambda_j}\right)^{-1/2}\{1 + o_P(1)\}\tilde{\boldsymbol{h}}_j = \left(1 + \frac{\delta}{\lambda_j}\right)^{-1/2}\{1 + o_P(1)\}(\boldsymbol{h}_j + \tilde{\boldsymbol{v}}_j). \tag{2.7}$$

The coefficient of $\boldsymbol{h}_j$ in $\hat{\boldsymbol{h}}_j$ depends on noise δ. The NR estimator $\tilde{\boldsymbol{h}}_j$ has an advantage over $\hat{\boldsymbol{h}}_j$ by applying the two steps described in (1.4) and (1.5) to the threshold estimation of PC directions.

We consider the following divergence conditions for d and n:

(⋆) $\log d/n = o(1)$ as $d \to \infty$ and $n \to \infty$.

Notably, (⋆) holds even when $d/n \to \infty$. Let $\tilde{\boldsymbol{h}}_j = (\tilde{h}_{j(1)}, \ldots, \tilde{h}_{j(d)})^T$ for all j.

Lemma 1. *Assume (A-i), (A-ii), (C-i), and (C-ii). Under (⋆), it holds for $j = 1, \ldots, m$ and $j' = 1, \ldots, d$ that*

$$\tilde{h}_{j(j')} = h_{j(j')} + O_P\Big\{(\lambda_j^{-1} n^{-1} \log d)^{1/2}\Big\} \quad \text{as } d \to \infty \text{ and } n \to \infty.$$

Thus, under the conditions in Lemma 1, we have consistency in the sense that $\tilde{h}_{j(j')} = h_{j(j')}\{1 + o_P(1)\}$ for j' satisfying $(n\lambda_j h_{j(j')}^2)^{-1} \log d = o(1)$. From (2.7), the elements in $\hat{\boldsymbol{h}}_j$ are not consistent unless $\delta/\lambda_j = o(1)$.

3. Automatic Sparse PCA Methodology

In this section, we propose a thresholded estimator of the PC directions by using the NR method. We emphasize that consistency properties of the SPCA methods heavily depend on threshold (tuning) values. To overcome the inconvenience, we propose an SPCA method from (2.5) and (2.6).

3.1. Thresholded estimator of PC directions using the NR method

For the PC direction $\boldsymbol{h}_j = (h_{j(1)}, \ldots, h_{j(d)})^T$, we arrange the elements $h_{j(1)}, \ldots, h_{j(d)}$ in the descending order as

$$|h_{oj(1)}| \geq \cdots \geq |h_{oj(d)}|. \tag{3.1}$$

Further, $\sum_{s=1}^d h_{oj(s)}^2 = 1$. Under (C-i), we assume the following condition for the PC direction as necessary:

(C-iii) For $j = 1, \ldots, m$, there exists an integer k_{j*} (which may depend on d), such that

$$\sum_{s=1}^{k_{j*}} h_{oj(s)}^2 \to 1 \ \text{ as } d \to \infty, \quad \liminf_{d\to\infty} \lambda_j h_{oj(k_{j*})}^2 > 0; \quad \text{and}$$
$$\limsup_{d\to\infty} \frac{|h_{oj(k_{j*}+1)}|}{|h_{oj(k_{j*})}|} < 1 \ \text{ when } k_{j*} \leq d-1.$$

Remark 2. When $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}_d$, $\boldsymbol{h}_1$ is $\mathbf{1}_d/d^{1/2}$, where $\boldsymbol{\Gamma}_d$ is given by (D.1) in Appendix D in the online supplementary material. Then, (C-iii) holds with $k_{1*} = d$ and $m = 1$.

Remark 6 in Section 5.2 and Appendix D present some models that satisfy (C-i) and (C-iii). From (G.15) in Appendix G in the online supplementary material, under (C-i) through (C-iii), we note that $k_{j*} \to \infty$ and $k_{j*}/\lambda_j \in (0, \infty)$ as $d \to \infty$ for $j = 1, \ldots, m$. Now, we consider removing $\tilde{\boldsymbol{v}}_j$ from $\tilde{\boldsymbol{h}}_j$ in (2.6) for each j (=

$1,\ldots,m$). Notably, $\|\tilde{\boldsymbol{v}}_j\| \approx \delta/\lambda_j$ from (2.5). For $\tilde{\boldsymbol{h}}_j = (\tilde{h}_{j(1)},\ldots,\tilde{h}_{j(d)})^T$ by the NR method, we arrange the elements $\tilde{h}_{j(1)},\ldots,\tilde{h}_{j(d)}$ in the descending order as $\tilde{h}_{oj(1)},\ldots,\tilde{h}_{oj(d)}$, where

$$|\tilde{h}_{oj(1)}| \geq \cdots \geq |\tilde{h}_{oj(d)}|.$$

Further, $\sum_{s=1}^d \tilde{h}^2_{oj(s)} = \|\tilde{\boldsymbol{h}}_j\|^2$. We consider the following thresholded estimator for an integer $k \in [1,d]$ as

$$\tilde{\boldsymbol{h}}_{oj}(k) = (\tilde{h}_{oj(1)},\ldots,\tilde{h}_{oj(k)},0,\ldots,0)^T$$

whose last $d-k$ elements are zero; that is, we replace $\tilde{h}_{oj(k+1)},\ldots,\tilde{h}_{oj(d)}$ in $(\tilde{h}_{oj(1)},\ldots,\tilde{h}_{oj(d)})^T$ with 0. Here, we provide the optimal integer k from (2.5) and (2.6). From $\|\tilde{\boldsymbol{h}}_j\| \geq 1$ w.p.1, there exists a unique integer $\tilde{k}_j \in [1,d]$ such that

$$\sum_{s=1}^{\tilde{k}_j-1} \tilde{h}^2_{oj(s)} < 1 \ \text{ and } \ \sum_{s=1}^{\tilde{k}_j} \tilde{h}^2_{oj(s)} \geq 1. \tag{3.2}$$

For a sequence $\{A_s\}$, we define $\sum_{s=1}^0 A_s = 0$ for convenience, so that $\tilde{k}_j = 1$ if $\tilde{h}^2_{oj(1)} \geq 1$. Notably, $\|\tilde{\boldsymbol{h}}_{oj}(\tilde{k}_j)\| \approx 1$ and $\sum_{s=\tilde{k}_j+1}^d \tilde{h}^2_{oj(s)} \approx \delta/\lambda_j$ for a sufficiently large d. Thus, we can remove $\tilde{\boldsymbol{v}}_j$ from $\tilde{\boldsymbol{h}}_j$ in (2.6) to obtain the following result. We write the elements of $\tilde{\boldsymbol{h}}_{oj}(\tilde{k}_j)$ as $\tilde{\boldsymbol{h}}_{oj}(\tilde{k}_j) = (\tilde{h}_{oj*(1)},\ldots,\tilde{h}_{oj*(d)})^T$. Further, we adjust the subscript j' of $\tilde{h}_{oj*(j')}$ as $\tilde{h}_{j*(j'')}$ corresponding to the order of elements in $\tilde{\boldsymbol{h}}_j$. Let $\tilde{\boldsymbol{h}}_{j*} = (\tilde{h}_{j*(1)},\ldots,\tilde{h}_{j*(d)})^T$ for all j. Let $\eta_j = \sum_{s=k_{j*}+1}^d h^2_{oj(s)}$ for $j = 1,\ldots,m$. Notably, $\eta_j \to 0$ as $d \to \infty$ in (C-iii).

Theorem 1. *Assume (A-i), (A-ii), and (C-i) through (C-iii). Under ($\star$), it holds for $j = 1,\ldots,m$ that $\|\tilde{\boldsymbol{h}}_{j*}\|^2 = 1 + O_P(\eta_j + n^{-1}) = 1 + o_P(1)$ and $\tilde{\boldsymbol{h}}^T_{j*}\boldsymbol{h}_j = 1 + O_P(\eta_j + n^{-1}) = 1 + o_P(1)$ as $d \to \infty$ and $n \to \infty$.*

From Theorem 1, we have the following result.

Corollary 1. *Assume (A-i), (A-ii), and (C-i) through (C-iii). Under ($\star$), it holds that*

$$\begin{aligned}
\|\tilde{\boldsymbol{h}}_{j*} - \boldsymbol{h}_j\|^2 &= O_P\big(\eta_j + n^{-1}\big) = o_P(1) \ \ \textit{for } j = 1,\ldots,m; \\
\tilde{\boldsymbol{h}}^T_{j*}\boldsymbol{h}_{j'} &= O_P\big(\eta_j^{1/2} + n^{-1/2}\min\{1, \lambda_j^{1/2}/\lambda_{j'}^{1/2}\}\big) = o_P(1) \ \ \textit{and} \\
\tilde{\boldsymbol{h}}^T_{j*}\tilde{\boldsymbol{h}}_{j'*} &= O_P\big(\eta_j^{1/2} + \eta_{j'}^{1/2} + n^{-1/2}\big) = o_P(1) \\
&\textit{when } m \geq 2 \ \ \textit{for } j, j' \leq m;\ j \neq j'
\end{aligned}$$

as $d \to \infty$ and $n \to \infty$.

Thus, $\tilde{\boldsymbol{h}}_{j*}$ has the aforementioned consistency properties even when $\delta/\lambda_j \to \infty$ without threshold (tuning) values such as ζ in (1.4). From Theorem 1, we have $\mathrm{Angle}(\tilde{\boldsymbol{h}}_{j*}, \boldsymbol{h}_j) = o_P(1)$ under the conditions in Corollary 1.

Proposition 3. *Assume (A-i) and (C-i). Then, it holds that* $\tilde{\boldsymbol{h}}_j^T \boldsymbol{h}_{j'} = O_P\{n^{-1/2}\min(1, \lambda_j^{1/2}/\lambda_{j'}^{1/2})\}$ *as* $d \to \infty$ *and* $n \to \infty$ *when* $m \geq 2$ *for* $j, j' \leq m;\ j \neq j'$.

From Proposition 3, $\tilde{\boldsymbol{h}}_{j*}^T \boldsymbol{h}_{j'}$ and $\tilde{\boldsymbol{h}}_j^T \boldsymbol{h}_{j'}$ are of the same order for $j, j' \leq m;\ j \neq j'$ if $\eta_j = O\{n^{-1/2}\min(1, \lambda_j^{1/2}/\lambda_{j'}^{1/2})\}$ as $d \to \infty$ and $n \to \infty$.

Remark 3. We assume that $|\tilde{h}_{oj(\tilde{k}_j)}| > |\tilde{h}_{oj(\tilde{k}_j+1)}|,\ j = 1, \ldots, m$ for the sake of simplicity. Then, we can simply obtain $\tilde{h}_{j*(j')}$ by

$$\tilde{h}_{j*(j')} = \begin{cases} \tilde{h}_{j(j')} & \text{if } |\tilde{h}_{j(j')}| \geq |\tilde{h}_{oj(\tilde{k}_j)}|\ , \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } j' = 1, \ldots, d.$$

3.2. Automatic SPCA

The computational cost of the SPCA method is high because it heavily depends on threshold (tuning) values determined by some cross-validation or information criteria. One can automatically yield the threshold estimation with a low computational cost because $\tilde{\boldsymbol{h}}_{j*}$ does not depend on any threshold (tuning) values such as ζ in (1.4).

We call the new PCA method that uses $\tilde{\lambda}_j$s and $\tilde{\boldsymbol{h}}_{j*}$s as the "automatic SPCA (A-SPCA)". Proposition 1 provides the details of $\tilde{\lambda}_j$. $\tilde{\lambda}_j$s have the consistency as "$\tilde{\lambda}_j/\lambda_j = 1 + o_P(1)$" even when $n/d \to 0$. The A-SPCA algorithm is given as

Automatic SPCA (A-SPCA)

Step 1. Set the dual-sample covariance matrix defined by $\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$.

Step 2. Find the eigenvalues $\hat{\lambda}_j$, and the eigenvectors $\hat{\boldsymbol{u}}_j$ of $\boldsymbol{S}_D$.

Step 3. Calculate $\tilde{\lambda}_j = \hat{\lambda}_j - \{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{s=1}^{j}\hat{\lambda}_s\}/(n-j-1)$ and $\tilde{\boldsymbol{h}}_j = \{(n-1)\tilde{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j$ for each j. Estimate the j-th eigenvalue by $\tilde{\lambda}_j$.

Step 4. Arrange the elements of $\tilde{\boldsymbol{h}}_j = (\tilde{h}_{j(1)}, \ldots, \tilde{h}_{j(d)})^T$ in the descending order as $|\tilde{h}_{oj(1)}| \geq \cdots \geq |\tilde{h}_{oj(d)}|$. Find the integer $\tilde{k}_j$ such that $\sum_{s=1}^{\tilde{k}_j-1}\tilde{h}_{oj(s)}^2 < 1$ and $\sum_{s=1}^{\tilde{k}_j}\tilde{h}_{oj(s)}^2 \geq 1$.

Step 5. Define $\tilde{\boldsymbol{h}}_{j*} = (\tilde{h}_{j*(1)}, \ldots, \tilde{h}_{j*(d)})^T$ by

$$\tilde{h}_{j*(j')} = \begin{cases} \tilde{h}_{j(j')} & \text{if } |\tilde{h}_{j(j')}| \geq |\tilde{h}_{oj(\tilde{k}_j)}|\ , \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } j' = 1, \ldots, d.$$

Estimate the j-th PC direction using $\tilde{\boldsymbol{h}}_{j*}$ for each j.

In Appendix A in the online supplementary material, we describe the application of A-SPCA to estimate $\boldsymbol{\Sigma}_1$. Appendix F presents an R code for A-SPCA.

Remark 4. Aoshima and Yata (2018) created a data-transformation method based on the strongly spiked eigenstructures and proposed a high-dimensional two-sample test using data transformation. Aoshima and Yata (2019) and Ishii, Yata and Aoshima (2022) proposed high-dimensional classifiers using data transformation. The key to these inferences is the estimation of the strongly spiked eigenstructures. In future, we apply A-SPCA to these inferences for high-dimensional data.

4. Simulation

In this section, we compare the performance of A-SPCA with the conventional PCA and TSPCA given by (1.5).

First, we set $d = 2^s,\ s = 6, \ldots, 12$ and $n = \lceil d^{1/2} \rceil$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. We generate $\boldsymbol{x}_i$, $i = 1, 2, \ldots$ independently from $N_d(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then, we consider the following two cases:

(S-i) $\boldsymbol{\Sigma}$ is given by $\lambda_1 = d^{2/3}$, $\lambda_2 = d^{1/2}$ and $\lambda_3 = \cdots = \lambda_d = 1$ together with $\boldsymbol{h}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{h}_2 = (0, 1, 0, \ldots, 0)^T$;

(S-ii) $\boldsymbol{\Sigma}$ is given by (D.2) in Appendix D in the online supplementary material with $\alpha = 0.5$, $\beta = 2$, $d_1 = \lceil d^{2/3} \rceil$, $d_2 = \lceil d^{1/2} \rceil$ and $\boldsymbol{\Omega}_{d-d_1-d_2} = \boldsymbol{I}_{d-d_1-d_2}$.

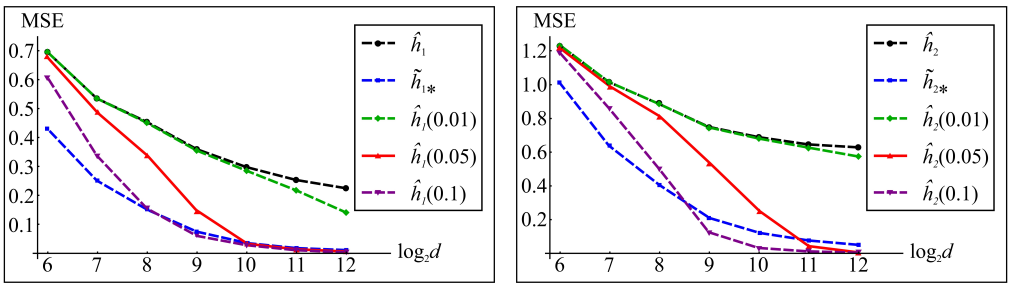
(S-i) satisfies (C-i) and (C-iii) when $m = 2$ and $k_{1*} = k_{2*} = 1$, and (S-ii) satisfies (C-i) and (C-iii) when $m = 2$, $k_{1*} = d_1$ and $k_{2*} = d_2$. (S-i) does not satisfy (C-ii) because $\sigma_{(1)} \geq \lambda_1 h_{1(1)}^2 \to \infty$. We considered (S-i) to handle a considerably sparse PC direction.

Next, we set $n = 10s,\ s = 2, \ldots, 10$ and $d = 1000$. We generated $\boldsymbol{x}_i$, $i = 1, 2, \ldots$ independently from the mixture model (5.4) with $g = 2$ and $\varepsilon_1 = \varepsilon_2 = 1/2$ as follows:

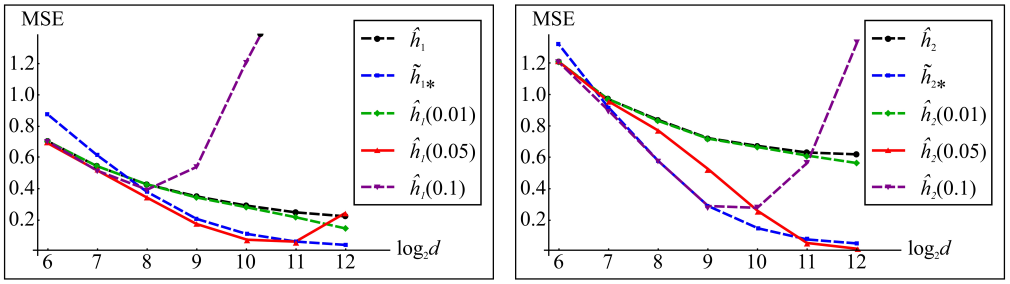
(S-iii) For $i = 1, 2$, $f_i(\boldsymbol{x};\ \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i)$ is the probability density function of $N_d(\boldsymbol{\mu}_i, \boldsymbol{\Psi}_i)$, where $\boldsymbol{\mu}_1 = (1, \ldots, 1, 0, \ldots, 0)^T$ the first $\lceil d^{2/3} \rceil$ elements are 1, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\Psi}_1 = (0.3^{|i-j|^{1/3}})$ and $\boldsymbol{\Psi}_2 = (0.4^{|i-j|^{1/3}})$.

Further, $\lambda_1 \approx d^{2/3}$ and $\boldsymbol{h}_1 \approx \boldsymbol{\mu}_1 / d^{1/3}$ in (S-iii). (S-iii) satisfies (C-i) and (C-iii) when $m = 1$. Details are provided in Remark 6 in Section 5.2. (S-iii) does not satisfy (A-i). Further details can be found in Section 4.1.1 in the work by Qiao et al. (2010).

Finally, we set $d = 2^s,\ s = 6, \ldots, 12$ and $n = \lceil d^{1/2} \rceil$. We consider a considerably non-sparse and non-Gaussian case for the first PC direction as follows:



(S-i) $N_d(\mathbf{0}, \mathbf{\Sigma})$, $d = 2^s$ $(s = 6, \ldots, 12)$, where $\mathbf{\Sigma}$ has $\lambda_1 = d^{2/3}$, $\lambda_2 = d^{1/2}$, and $\lambda_3 = \cdots = \lambda_d = 1$ together with $\boldsymbol{h}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{h}_2 = (0, 1, 0, \ldots, 0)^T$.



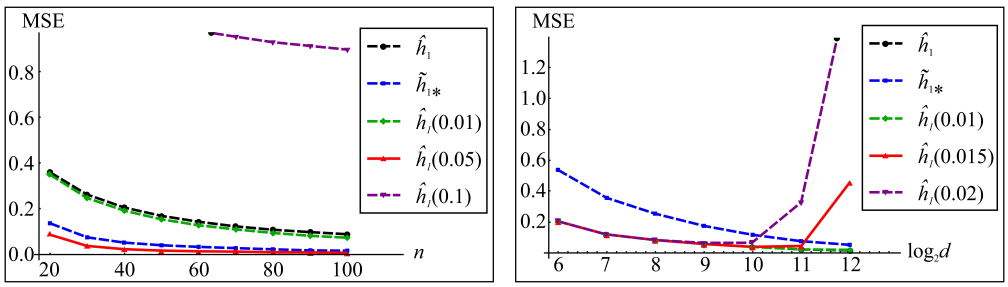
(S-ii) $N_d(\mathbf{0}, \mathbf{\Sigma})$, $d = 2^s$ $(s = 6, \ldots, 12)$, where $\mathbf{\Sigma}$ has $\lambda_1 \approx d^{2/3}$ and $\lambda_2 \approx d^{1/2}$ together with $\boldsymbol{h}_1 = (1, \ldots, 1, \ldots, 0)^T$ whose $\lceil d^{2/3} \rceil$ elements are 1 and $\boldsymbol{h}_2 = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)^T$, whose $\lceil d^{1/2} \rceil$ elements are 1.

Figure 2. Average mean-squared errors for PC directions in (S-i) and (S-ii).

(S-iv) We generated $\boldsymbol{x}_i = \boldsymbol{H}\mathbf{\Lambda}^{1/2}\boldsymbol{z}_i$, $i = 1, 2, \ldots$ independently from $z_{ij} = (y_{ij} - 5)/10^{1/2}$ $(j = 1, \ldots, d)$ where y_{ij}s are i.i.d. as the chi-squared distribution with 5 degrees of freedom. We set $\mathbf{\Sigma} = \mathbf{\Gamma}_d$ where $\alpha = 0.5$ and $\beta = 1$, and $\mathbf{\Gamma}_d$ is given by (D.1) in Appendix D.

$\lambda_1 = 0.5d + 0.5$ and $\boldsymbol{h}_1 = d^{-1/2}\mathbf{1}_d$ in (S-iv). Further, (S-iv) satisfies (A-i), (C-i), and (C-iii) when $m = 1$ and $k_{1*} = d$. However, (S-iv) does not satisfy (A-ii). Here, $\boldsymbol{h}_1$ appears to be a non-sparse vector in the sense that all elements of $\boldsymbol{h}_1$ are nonzero.

We set $\zeta = 0.01$, 0.05 and 0.1 for TSPCA using (1.5) in (S-i) to (S-iii). Further, $\zeta = 0.01$ is a soft threshold and $\zeta = 0.1$ is a hard threshold. In (S-iv), we set $\zeta = 0.01$, 0.015 and 0.02 for TSPCA because all elements of $\boldsymbol{h}_1$ are nonzero and small. Thus, $\hat{\boldsymbol{h}}_{j(\zeta)}$ with a hard threshold results in a considerably poor performance in (S-iv). We considered the mean-squared error $\mathrm{MSE}(\hat{\boldsymbol{h}}_j) = \|\hat{\boldsymbol{h}}_j - \boldsymbol{h}_j\|^2$ $(j = 1, \ldots, m)$ and took the average of its outcomes from 2,000 independent replications. Similar procedures were performed for $\mathrm{MSE}(\tilde{\boldsymbol{h}}_{j*})$ and $\mathrm{MSE}(\hat{\boldsymbol{h}}_{j(\zeta)})$. We summarized the results in Figures 2 and 3.



(S-iii) (S-iv)

(S-iii) Mixture model (5.4) with $g = 2$, $d = 1000$, $n = 10s$ ($s = 2, \ldots, 10$), $\lambda_1 \approx d^{2/3}$ and $\boldsymbol{h}_1 \approx d^{-1/3}(1, \ldots, 1, 0, \ldots, 0)^T$ whose first $\lceil d^{2/3} \rceil$ elements are $d^{-2/3}$. (S-iv) $\boldsymbol{x}_i = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}_i$; $z_{ij} = (y_{ij} - 5)/10^{1/2}$ ($j = 1, \ldots, d$) in which y_{ij}s are i.i.d. as the chi-squared distribution with 5 degrees of freedom, $d = 2^s$ ($s = 6, \ldots, 12$), $\lambda_1 = 0.5d + 0.5$, and $\boldsymbol{h}_1 = d^{-1/2}\boldsymbol{1}_d$.

Figure 3. Average mean-squared errors for the first PC direction in (S-iii) and (S-iv). In the left panel, $\mathrm{MSE}(\hat{\boldsymbol{h}}_1(0.1))$ is too high to describe when n is small.

As expected, we observed that A-SPCA achieved preferable performances in (S-i) to (S-iii). For (S-iv), the conventional PCA or TSPCA with $\zeta = 0.01$ performed better than A-SPCA because all elements of $\boldsymbol{h}_1$ are nonzero. In addition, λ_1 is considerably large because $\lambda_1 = O(d)$. $\hat{\lambda}_1$ is consistent in the sense that $\hat{\lambda}_1/\lambda_1 = 1 + o_P(1)$ (see Appendix E in the online supplementary material). However, A-SPCA performed preferably as d increased, even for the non-sparse case. TSPCA with $\zeta = 0.015$ and $\zeta = 0.02$ exhibited poor performances when d is large. This is because all elements of $\boldsymbol{h}_1 = (1/d^{1/2}, \ldots, 1/d^{1/2})^T$ in (S-iv) become close to zero with an increase in d.

Throughout the simulations, TSPCA depended heavily on the threshold value. In contrast, A-SPCA performed preferably without using any threshold values.

5. Shrinkage PC Directions and its Application to Clustering

We consider the shrinkage PC directions using A-SPCA, and we apply them to clustering. Figure 1 shows that the TSPCA yields a preferable performance even though it may not hold the consistency as in Corollary 1. Further, we show that PCA in the shrinkage PC direction is effective for clustering high-dimensional data. We emphasize that the shrinkage PC directions depend on a parameter unlike in A-SPCA.

5.1. Shrinkage PC directions

For a given constant $\omega_j \in (0, 1]$ for each j ($= 1, \ldots, m$), we consider the PC shrinkage directions with a cumulative contribution ratio greater than or equal to ω_j. For each j ($= 1, \ldots, m$), there exists a unique integer $k_{j\omega} \in [1, d]$ such that

$$\sum_{s=1}^{k_{j\omega}-1} h_{oj(s)}^2 < \omega_j \quad \text{and} \quad \sum_{s=1}^{k_{j\omega}} h_{oj(s)}^2 \geq \omega_j, \tag{5.1}$$

where $h_{oj(s)}$ are given in (3.1). We assume that $|h_{oj(k_{j\omega})}| > |h_{oj(k_{j\omega}+1)}|$ for simplicity. We define

$$h_{j\omega(j')} = \begin{cases} h_{j(j')} & \text{if } |h_{j(j')}| \geq |h_{oj(k_{j\omega})}| \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } j' = 1, \ldots, d.$$

Let $\boldsymbol{h}_{j\omega} = (h_{j\omega(1)}, \ldots, h_{j\omega(d)})^T$ for $j = 1, \ldots, m$. Further, $\|\boldsymbol{h}_{j\omega}\|^2 = \sum_{s=1}^{k_{j\omega}} h_{oj(s)}^2$ $(\geq \omega_j)$ if $|h_{oj(k_{j\omega})}| > |h_{oj(k_{j\omega}+1)}|$. We seek to estimate the shrinkage PC direction $\boldsymbol{h}_{j\omega}$.

As in (3.2), there exists a unique integer $\tilde{k}_{j\omega} \in [1, d]$ such that

$$\sum_{s=1}^{\tilde{k}_{j\omega}-1} \tilde{h}_{oj(s)}^2 < \omega_j \quad \text{and} \quad \sum_{s=1}^{\tilde{k}_{j\omega}} \tilde{h}_{oj(s)}^2 \geq \omega_j. \tag{5.2}$$

We assume that $|\tilde{h}_{oj(\tilde{k}_{j\omega})}| > |\tilde{h}_{oj(\tilde{k}_{j\omega}+1)}|$ for simplicity. We define that

$$\tilde{h}_{j\omega(j')} = \begin{cases} \tilde{h}_{j(j')} & \text{if } |\tilde{h}_{j(j')}| \geq |\tilde{h}_{oj(\tilde{k}_{j\omega})}| \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } j' = 1, \ldots, d.$$

Let $\tilde{\boldsymbol{h}}_{j\omega} = (\tilde{h}_{j\omega(1)}, \ldots, \tilde{h}_{j\omega(d)})^T$ for $j = 1, \ldots, m$. Notably, $\tilde{\boldsymbol{h}}_{j\omega} = \tilde{\boldsymbol{h}}_{j*}$ when $\omega_j = 1$. Under (C-i), we assume the following conditions:

(C-iii') For some fixed integers r_j (≥ 0), $\limsup_{d\to\infty} |h_{oj(k_{j\omega}+r_j+1)}|/|h_{oj(k_{j\omega}+r_j)}| < 1$ and $\liminf_{d\to\infty} \lambda_j h_{oj(k_{j\omega}+r_j)}^2 > 0$, and $\omega_j\lambda_j \to \infty$ as $d \to \infty$ for $j = 1, \ldots, m$.

From (G.47) in Appendix G in the online supplementary material, under (C-i), (C-ii), and (C-iii'), we note that $k_{j\omega} \to \infty$ and $k_{j\omega}/(\omega_j\lambda_j) \in (0, \infty)$ as $d \to \infty$ for $j = 1, \ldots, m$.

Theorem 2. *Assume (A-i), (A-ii), (C-i), (C-ii), and (C-iii'). Under (⋆), it holds for $j = 1, \ldots, m$ that $\|\tilde{\boldsymbol{h}}_{j\omega} - \boldsymbol{h}_{j\omega}\|^2 = O_P\{\omega_j(k_{j\omega}^{-1} + n^{-1/2})\} = o_P(\omega_j)$ as $d \to \infty$ and $n \to \infty$.*

Thus, $\tilde{\boldsymbol{h}}_{j\omega}$ is consistent even when $\delta/\lambda_j \to \infty$. However, we cannot construct a consistent estimator of $\boldsymbol{h}_{j\omega}$ using $\hat{\boldsymbol{h}}_j$ unless $\delta/\lambda_j = o(1)$. The reason is explained in (2.7).

We propose shrinkage PC (SH-PC) scores using the shrinkage PC directions as

$$(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \tilde{\boldsymbol{h}}_{j\omega}, \quad i = 1, \ldots, n \tag{5.3}$$

for each j. In Section 6, we investigated the performance of SH-PC scores for

real datasets.

Remark 5. One can use a normalized shrinkage PC direction $\tilde{\boldsymbol{h}}_{j\omega}/\|\tilde{\boldsymbol{h}}_{j\omega}\|$ in (5.3).

5.2. Application to clustering

For the population distribution of $\boldsymbol{x}_i$, we consider a g (≥ 2)-class mixture model whose probability density function is given by

$$f(\boldsymbol{x}_i) = \sum_{s=1}^{g} \varepsilon_s f_s(\boldsymbol{x}_i;\ \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s), \tag{5.4}$$

where $\varepsilon_s \in (0,1)$, $s = 1,\ldots,g$, $\sum_{s=1}^{g}\varepsilon_s = 1$ and $f_s(\boldsymbol{x}_i;\ \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s)$ represents the probability density function of a d-variate population Π_s with mean $\boldsymbol{\mu}_s$ and covariance matrix $\boldsymbol{\Psi}_s$. $E(\boldsymbol{x}_i)$ $(= \boldsymbol{\mu})$ $= \sum_{s=1}^{g}\varepsilon_s\boldsymbol{\mu}_s$ and $\mathrm{Var}(\boldsymbol{x}_i) = \sum_{s=1}^{g-1}\sum_{s'=s+1}^{g}\varepsilon_s\varepsilon_{s'}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{s'})(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{s'})^T + \sum_{s=1}^{g}\varepsilon_s\boldsymbol{\Psi}_s$ $(= \boldsymbol{\Sigma})$.

Two-class mixture model. We assume $g = 2$. Let $\boldsymbol{\mu}_{12} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\Delta_{12} = \|\boldsymbol{\mu}_{12}\|^2$. Then, the covariance matrix of the mixture model is given by $\boldsymbol{\Sigma} = \varepsilon_1\varepsilon_2\boldsymbol{\mu}_{12}\boldsymbol{\mu}_{12}^T + \varepsilon_1\boldsymbol{\Psi}_1 + \varepsilon_2\boldsymbol{\Psi}_2$. We assume $\boldsymbol{h}_1^T\boldsymbol{\mu}_{12} \geq 0$ without loss of generality. If

$$\frac{\lambda_{\max}(\boldsymbol{\Psi}_s)}{\Delta_{12}} \to 0 \ \text{ as } d \to \infty \text{ for } s = 1,2, \tag{5.5}$$

it holds that

$$\frac{\lambda_1}{\varepsilon_1\varepsilon_2\Delta_{12}} \to 1 \ \text{ and } \ \mathrm{Angle}(\boldsymbol{h}_1, \boldsymbol{\mu}_{12}) \to 0 \ \text{ as } d \to \infty, \tag{5.6}$$

where $\lambda_{\max}(\boldsymbol{M})$ denotes the largest eigenvalue of any positive-semidefinite matrix $\boldsymbol{M}$. Thus, we can obtain a threshold estimation of $\boldsymbol{\mu}_{12}/\Delta_{12}^{1/2}$ using A-SPCA. Furthermore, for the first (true) PC score, it follows that

$$\boldsymbol{h}_1^T(\boldsymbol{x}_i - \boldsymbol{\mu}) = \begin{cases} \lambda_1^{1/2}\{(\varepsilon_2/\varepsilon_1)^{1/2} + o_P(1)\} & \text{when } \boldsymbol{x}_i \in \Pi_1, \\ -\lambda_1^{1/2}\{(\varepsilon_1/\varepsilon_2)^{1/2} + o_P(1)\} & \text{when } \boldsymbol{x}_i \in \Pi_2 \end{cases} \tag{5.7}$$

as $d \to \infty$ under (5.5). Hence, one can cluster $\boldsymbol{x}_i$s into two groups based on the sign of the estimated first PC scores, as shown in Figure 1(i). Sections 2 and 3 in the work by Yata and Aoshima (2020) provide further details on (5.6) and (5.7).

Let $\boldsymbol{\mu}_{12} = (\mu_{(1)},\ldots,\mu_{(d)})^T$. We assume $|\mu_{(1)}| \geq \cdots \geq |\mu_{(d)}|$ without loss of generality.

Remark 6. If it holds that

$$\mu_{(k+1)} = \cdots = \mu_{(d)} = 0, \ \ \liminf_{d\to\infty}|\mu_{(k)}| > 0 \ \text{ and } \ k \geq d^{1/2} \text{ for some integer } k,$$

(C-i) and (C-iii) with $m = 1$ are satisfied when $\max_{s=1,2} \mathrm{tr}(\boldsymbol{\Psi}_s^2) = O(d)$ and $\boldsymbol{\Psi}_s$s have NSSE models.

For an integer k, we write $\boldsymbol{x}_{i(k)} = (x_{i(1)}, \ldots, x_{i(k)})^T$ for all i, $\mathrm{Var}(\boldsymbol{x}_{i(k)}) = \boldsymbol{\Sigma}_{(k)}$, and $\mathrm{Var}(\boldsymbol{x}_{i(k)}) = \boldsymbol{\Psi}_{s(k)}$ when $\boldsymbol{x}_i \in \Pi_s$. Let $\boldsymbol{h}_{1(k)} = (h_{1(1)}, \ldots, h_{1(k)})^T$, $\boldsymbol{\mu}_{12(k)} = (\mu_{(1)}, \ldots, \mu_{(k)})^T$, and $\Delta_{12(k)} = \|\boldsymbol{\mu}_{12(k)}\|^2$. As in (5.6) and (5.7), under the condition that

$$\frac{\lambda_{\max}(\boldsymbol{\Psi}_{s(k)})}{\Delta_{12(k)}} \to 0 \quad \text{as } d \to \infty \text{ for } s = 1, 2, \tag{5.8}$$

the following holds: $\mathrm{Angle}(\boldsymbol{h}_{1(k)}, \boldsymbol{\mu}_{12(k)}) \to 0$ as $d \to \infty$. Furthermore, for the first PC score, it follows that as $d \to \infty$

$$\boldsymbol{h}_{1(k)}^T(\boldsymbol{x}_i - \boldsymbol{\mu}) = \begin{cases} \{\lambda_{\max}(\boldsymbol{\Sigma}_{(k)})\}^{1/2}\{(\varepsilon_2/\varepsilon_1)^{1/2} + o_P(1)\} & \text{when } \boldsymbol{x}_i \in \Pi_1, \\ -\{\lambda_{\max}(\boldsymbol{\Sigma}_{(k)})\}^{1/2}\{(\varepsilon_1/\varepsilon_2)^{1/2} + o_P(1)\} & \text{when } \boldsymbol{x}_i \in \Pi_2. \end{cases}$$

Thus, the shrinkage (true) scores were consistent with those under (5.8). If

$$\liminf_{d\to\infty} \frac{\Delta_{12(k)}}{\Delta_{12}} > 0 \quad \text{and} \quad \lambda_{\max}(\boldsymbol{\Psi}_{s(k)}) = o\{\lambda_{\max}(\boldsymbol{\Psi}_s)\}, \tag{5.9}$$

(5.8) is milder than that in (5.5). The second condition in (5.9) often holds when $k \ll d$. Further, from (5.6), it holds that $\boldsymbol{h}_{1(k)} \approx \boldsymbol{\mu}_{12(k)}/\Delta_{12}^{1/2}$, so that

$$\frac{\Delta_{12(k)}}{\Delta_{12}} \approx \|\boldsymbol{h}_{1(k)}\|^2 = \sum_{s=1}^{k} h_{oj(s)}^2.$$

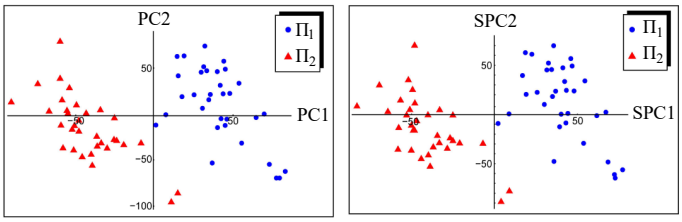
On the basis of (5.1) and (5.9), if $\liminf_{d\to\infty} \omega_j > 0$, the SH-PC scores were consistent under milder conditions than those for the PC scores. The SH-PC scores by $\tilde{\boldsymbol{h}}_{j\omega}$ effectively cluster $\boldsymbol{x}_i$s into two groups (see Fig. 1(iii) and Sec. 6.1).

g (≥ 3)-class mixture model. Let $\boldsymbol{\mu}_{23} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$. When $g = 3$, from Theorem 2 and Lemma 2 in Yata and Aoshima (2020), under certain regularity conditions, it holds that $\mathrm{Angle}(\boldsymbol{h}_1, \boldsymbol{\mu}_{12}) \to 0$ and $\mathrm{Angle}(\boldsymbol{h}_2, \boldsymbol{\mu}_{23}) \to 0$ as $d \to \infty$. Furthermore, for (true) PC scores, it follows that

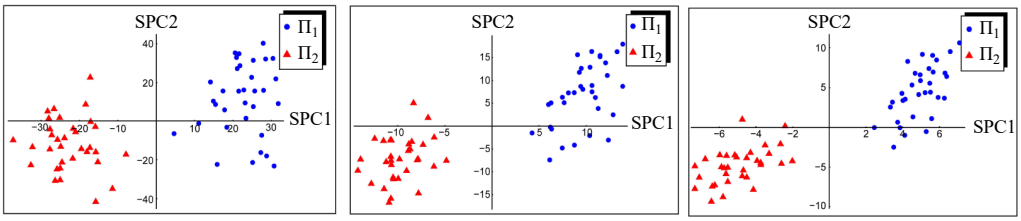
$$\boldsymbol{h}_1^T(\boldsymbol{x}_i - \boldsymbol{\mu}) = \begin{cases} \lambda_1^{1/2}\big[\{(1-\varepsilon_1)/\varepsilon_1\}^{1/2} + o_P(1)\big] & \text{when } \boldsymbol{x}_i \in \Pi_1, \\ -\lambda_1^{1/2}\big[\{\varepsilon_1/(1-\varepsilon_1)\}^{1/2} + o_P(1)\big] & \text{otherwise} \end{cases} \tag{5.10}$$

and

$$\boldsymbol{h}_2^T(\boldsymbol{x}_i - \boldsymbol{\mu}) = \begin{cases} o_P(\lambda_2^{1/2}) & \text{when } \boldsymbol{x}_i \in \Pi_1, \\ \lambda_2^{1/2}\big([\varepsilon_3/\{\varepsilon_2(1-\varepsilon_1)\}]^{1/2} + o_P(1)\big) & \text{when } \boldsymbol{x}_i \in \Pi_2, \\ -\lambda_2^{1/2}\big([\varepsilon_2/\{\varepsilon_3(1-\varepsilon_1)\}]^{1/2} + o_P(1)\big) & \text{when } \boldsymbol{x}_i \in \Pi_3, \end{cases} \tag{5.11}$$



(i) Conventional PCA (ii) A-SPCA



(iii) SH-PCs with $\omega = 0.5$ (iv) SH-PCs with $\omega = 0.2$ (v) SH-PCs with $\omega = 0.1$

Figure 4. Scatter plots of the first two PC scores for the dataset comprising Π_1: B-cell with 33 samples and Π_2: T-cell with 33 samples.

for $i = 1, \ldots, n$. Thus, as in the case of $g = 2$, the SH-PC scores by $\tilde{\boldsymbol{h}}_{j\omega}$ with $\liminf_{d\to\infty} \omega_j > 0$ can effectively cluster $\boldsymbol{x}_i$s into three groups (see Section 6.2).

When $g \geq 4$, Yata and Aoshima (2020) provided the consistency properties of the PC scores.

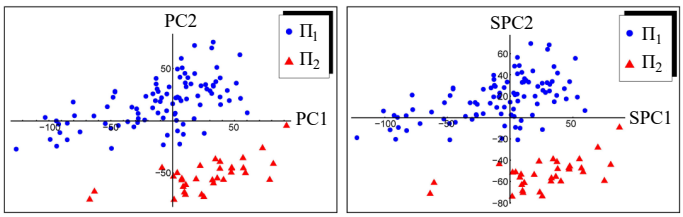
6. Example: Clustering

We assess the performance of clustering based on SH-PC scores using (5.3). We choose $\omega_1 = \omega_2$ $(= \omega,$ say) in (5.2) as $\omega = 0.1,\ 0.2$ and 0.5.

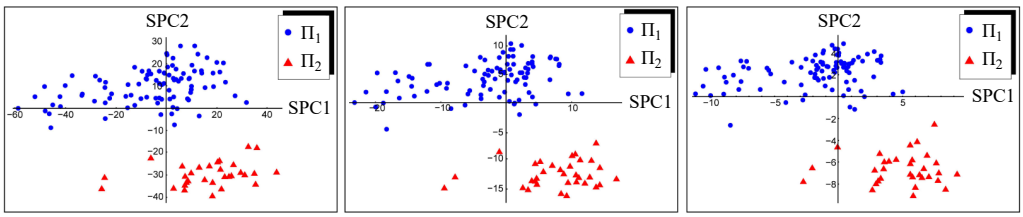
6.1. Two-class mixture model

We used microarray data provided by Chiaretti et al. (2004) with $d = 12625$ (see Fig. 1). We consider a dataset consisting of 33 samples from Π_1: B-cell and 33 samples from Π_2: T-cell. For the mixed 66 samples, we calculated the first two PC scores using PCA, A-SPCA, and SH-PC with $\omega = 0.1,\ 0.2$, and 0.5. We have that $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (17, 50)$, $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (52, 134)$ and $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (381, 640)$ for $\omega = 0.1,\ 0.2$ and 0.5, respectively. Figure 4 depicts the scatter plots of the PC scores. We observed that the 66 samples were perfectly classified into two groups based on the sign of the first PC scores even when the cumulative contribution ratio ω is as small as $\omega = 0.1$. These data can be clustered using the SH-PC score with only 17 significant variables. Details are provided in Section 5.2.

Next, we consider an unbalanced case. The B-cells originally contained 95 samples. We set Π_1: B-cell with 95 samples and Π_2: T-cell with 33 samples. As in



(i) Conventional PCA (ii) A-SPCA



(iii) SH-PCs with $\omega = 0.5$ (iv) SH-PCs with $\omega = 0.2$ (v) SH-PCs with $\omega = 0.1$

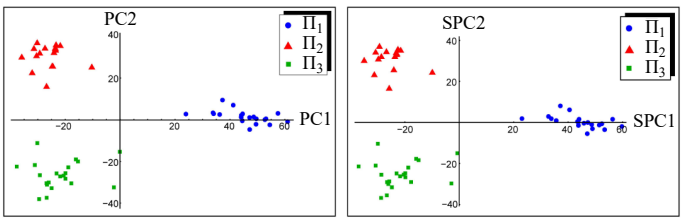
Figure 5. Scatter plots of the first two PC scores for the dataset that comprises Π_1: B-cell with 95 samples and Π_2: T-cell with 33 samples.

Figure 4, we present the scatter plots of the PC scores in Figure 5. The following hold: $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (88, 13)$, $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (229, 35)$, and $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (977, 223)$ for $\omega = 0.1,\ 0.2$, and 0.5, respectively. The 128 samples were effectively classified into two groups based on the sign of the second SH-PC score; this figure is remarkably different from Figure 4. Further, $\varepsilon_1\varepsilon_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ becomes small in such an imbalanced case, and therefore, $\lambda_{\max}(\boldsymbol{\Sigma}_1)$ or $\lambda_{\max}(\boldsymbol{\Sigma}_2)$ is probably the largest eigenvalue of $\boldsymbol{\Sigma}$ because $\boldsymbol{\Sigma} = \varepsilon_1\varepsilon_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T + \varepsilon_1\boldsymbol{\Sigma}_1 + \varepsilon_2\boldsymbol{\Sigma}_2$. The second PC direction is probably $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\Delta^{1/2}$. Thus, 128 samples were classified into two groups according to the sign of the second SH-PC scores. Section 4.3 in the work by Yata and Aoshima (2020) for further details.

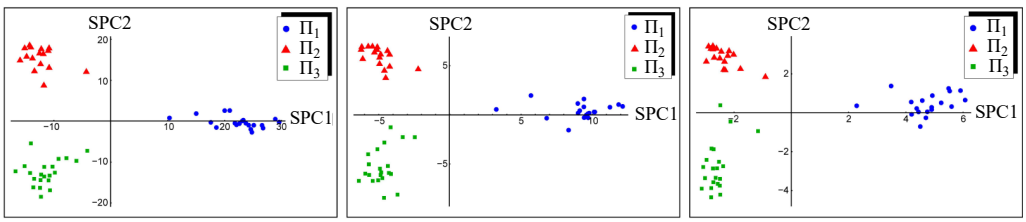
6.2. Three-class mixture model

We analyzed microarray data from Bhattacharjee et al. (2001) in which the dataset comprised five lung carcinoma types with $d = 3312$. We used only three classes: Π_1 : pulmonary carcinoids with 20 samples; Π_2 : normal lung with 17 samples; and Π_3 : squamous cell lung carcinoma with 21 samples. As in Figure 4, we present the scatter plots of the PC scores in Figure 6. The following hold: $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (17, 16)$, $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (44, 41)$, and $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (195, 171)$ for $\omega = 0.1,\ 0.2$ and 0.5, respectively.

We observed that all samples were effectively classified into three groups based on the first two PC scores. Further details are provided in (5.10) and (5.11). Figure 6(v) shows that these data can be clustered using the SH-PC score with $(\tilde{k}_{1\omega}, \tilde{k}_{2\omega}) = (17, 16)$. Here, the 17 variables in $\boldsymbol{h}_{1\omega}$ and the 16 variables in



(i) Conventional PCA (ii) A-SPCA



(iii) SH-PCs with $\omega = 0.5$ (iv) SH-PCs with $\omega = 0.2$ (v) SH-PCs with $\omega = 0.1$

Figure 6. Scatter plots of the first two PC scores for the dataset that comprises Π_1 : pulmonary carcinoids with 20 samples; Π_2 : normal lung with 17 samples; and Π_3 : squamous cell lung carcinomas with 21 samples.

$\boldsymbol{h}_{2\omega}$ were all different. Therefore, these data can be clustered using the first two SH-PC scores with only 33 significant variables.

7. Conclusion

In this study, we investigated TSPCA in high-dimensional settings. We proposed a new TSPCA method called automatic SPCA (A-SPCA) and demonstrated that it exhibits the consistency property under mild conditions free from any threshold values. Therefore, we can quickly obtain a more accurate result at a lower computational cost. Further, we proposed shrinkage PC directions and applied them to the clustering. We demonstrated the performance of clustering based on the shrinkage PC directions. We demonstrated that the datasets could be clustered using a set of significant variables.

Supplementary Material

We give an estimate of the intrinsic part in $\boldsymbol{\Sigma}$ using A-SPCA, examples of the strongly spiked eigenstructures, asymptotic results for A-SPCA under a milder assumption than (A-ii), comparisons between TSPCA and RSPCA, asymptotic properties of the conventional PCA, an R code for A-SPCA and proofs of the theoretical results in the online Supplementary Material.

Acknowledgments

We are very grateful to the associate editor and the reviewers for their constructive comments. The research of the first author was partially supported by Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS), under Contract Number 22K03412. The research of the second author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 20H00576 and 22K19769.

References

Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics* **34**, 122–148.

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's Special Invited Paper)* **30**, 356–399.

Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica* **28**, 43–62.

Aoshima, M. and Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics* **71**, 473–503.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

Bai, Z. and Ding, X. (2012). Estimation of spiked eigenvalues in spiked models. *Random Matrices: Theory and Applications* **1**, 1150011.

Baik, J. and Silverstein, J.W. (2006) Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382–1408.

Bhattacharjee A., Richards W. G., Staunton, J., Li, C., Monti, S., Vasa, P. et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. In *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13790–13795.

Cai. T. T., Han, X. and Pan, G. (2020). Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *The Annals of Statistics* **48**, 1255–1280.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F. et al. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771–2778.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **75**, 603–680.

Ishii, A., Yata, K. and Aoshima, M. (2022). Geometric classifiers for high-dimensional noisy data. *Journal of Multivariate Analysis (Editor's Invited Paper)* **188**, 104850.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29**, 295–327.

Johnstone, I. and Lu, Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **104**, 682–693.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics* **37**, 4104–4130.

Lee, S., Huang, J. Z. and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics* **4**, 1579–1601.

Lee, S., Zou, F. and Wright F.A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *The Annals of Statistics* **38**, 3605–3629.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics* **41**, 772–801.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168**, 244–258.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* **17**, 1617–1642.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference* **150**, 1–29.

Paul, D. and Johnstone, I. M. (2007). Augmented sparse principal component analysis for high dimensional data. Technical Report. UC Davis, Davis.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J. and Marron, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **105**, 401–414.

Shen, H. and Huang, J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation *Journal of Multivariate Analysis* **99**, 1015–1034.

Shen, D., Shen, H. and Marron, J.S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis* **115**, 317–333.

Shen, D., Shen, H., Zhu, H. and Marron, J.S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica* **26**, 1747–1770.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics* **45**, 1342–1374.

Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. *Communications in Statistics - Theory and Methods, Special Issue Honoring Zacks, S. (Edited by N. Mukhopadhyay)* **38**, 2634–2652.

Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* **105**, 193–215.

Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis* **122**, 334–354.

Yata, K. and Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics* **47**, 899–921.

Zou, H. and Hastie, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286.

(Received September 2022; accepted July 2023)