Received: 29 March 2018 | Revised: 17 May 2019 | Accepted: 21 September 2019

DOI: 10.1111/sjos.12432

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Geometric consistency of principal component scores for high-dimensional mixture models and its application

Kazuyoshi Yata | **Makoto Aoshima**

Institute of Mathematics, University of Tsukuba

Correspondence
Makoto Aoshima, Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan.
Email: aoshima@math.tsukuba.ac.jp

Funding information
Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS), 26800078; Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, 15H01678, 17K19956

Abstract

In this article, we consider clustering based on principal component analysis (PCA) for high-dimensional mixture models. We present theoretical reasons why PCA is effective for clustering high-dimensional data. First, we derive a geometric representation of high-dimension, low-sample-size (HDLSS) data taken from a two-class mixture model. With the help of the geometric representation, we give geometric consistency properties of sample principal component scores in the HDLSS context. We develop ideas of the geometric representation and provide geometric consistency properties for multiclass mixture models. We show that PCA can cluster HDLSS data under certain conditions in a surprisingly explicit way. Finally, we demonstrate the performance of the clustering using gene expression datasets.

KEYWORDS

clustering, geometric representation, HDLSS, microarray, PCA, PC score

1 | INTRODUCTION

High-dimension, low-sample-size (HDLSS) data situations occur in many areas of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. In recent years, substantial work has been done on HDLSS asymptotic theory, where the sample size n is fixed or $n/d \to 0$ as the data dimension $d \to \infty$. Hall, Marron, and Neeman (2005), Ahn, Marron, Muller, and Chi (2007), Yata and Aoshima (2012), and Lv (2013) explored several types of geometric representations of HDLSS data. Jung and Marron (2009) showed inconsistency properties of the sample eigenvalues and eigenvectors in the HDLSS context. Yata and Aoshima (2012, 2013) developed the noise-reduction methodology to provide consistent estimators of both the eigenvalues and eigenvectors together with principal component (PC) scores in the HDLSS context. Shen, Shen, Zhu, and Marron (2016) and Hellton and Thoresen (2017) also provided several asymptotic properties of the sample PC scores in the HDLSS context.

The HDLSS asymptotic theory was created under the assumption of either the population distribution is Gaussian or the random variables in a sphered data matrix have a ρ-mixing dependency. However, Yata and Aoshima (2010) developed an HDLSS asymptotic theory without such assumptions. Moreover, they created a new principal component analysis (PCA) called the cross-data-matrix methodology that is applicable to construct an unbiased estimator in HDLSS nonparametric settings. Based on the cross-data-matrix methodology, Aoshima and Yata (2011) developed a variety of inference for HDLSS data such as given-bandwidth confidence region, two-sample test, classification, variable selection, regression, pathway analysis, and so on (see Aoshima et al., 2018 for the review).

PCA is an important visualization and dimension reduction technique for high-dimensional data. Furthermore, PCA is quite popular for clustering high-dimensional data (see section 9.2 in Jolliffe, 2002 for details). For clustering HDLSS gene expression data, see Armstrong et al. (2002) and Pomeroy et al. (2002). Liu, Hayes, Nobel, and Marron (2008) and Ahn, Lee, and Yoon (2012) gave binary split-type clustering methods for HDLSS data. Borysov, Hannig, and Marron (2014) considered hierarchical clustering for high-dimensional data. Li and Yao (2018) considered a model-based clustering for a high-dimensional mixture. Given this background, we decided to focus on high-dimensional structures of multiclass mixture models via PCA. In this article, we consider asymptotic properties of PC scores for high-dimensional mixture models to apply for cluster analysis in HDLSS settings. The main contribution of this article is that we give theoretical reasons why PCA is effective for clustering HDLSS data.

Suppose there are independent and d-variate populations, $\Pi_i, i = 1, \dots, k$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown (positive-semidefinite) covariance matrix $\boldsymbol{\Sigma}_i$ for each i. We consider a mixture model to classify a dataset into k (≥ 2) groups. We assume that any sample is taken with mixing proportions ε_is from Π_is, where $\varepsilon_i \in (0, 1)$ and $\sum_{i=1}^{k} \varepsilon_i = 1$ but the label of the population is missing. We assume that k and ε_is are independent of d. We consider a mixture model whose probability density function (or probability function) is given by

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} \varepsilon_i \pi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $\pi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is a d-dimensional probability density function (or probability function) of Π_i having a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Suppose we have a $d \times n$ data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_j, j = 1, \dots, n$, are independently taken from Equation (1). We assume $n \geq k$. Let

$$n_i = \#\{j | \boldsymbol{x}_j \in \Pi_i \text{ for } j = 1, \dots, n\} \quad \text{and} \quad \eta_i = n_i/n \text{ for } i = 1, \dots, k,$$

where $\#A$ denotes the number of elements in a set A. We assume that n and n_is are independent of d. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the mean vector and the covariance matrix of Equation (1), respectively. Then, we have that

$$\boldsymbol{\mu} = \sum_{i=1}^{k} \varepsilon_i \boldsymbol{\mu}_i \quad \text{and} \quad \boldsymbol{\Sigma} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \varepsilon_i \varepsilon_j (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} + \sum_{i=1}^{k} \varepsilon_i \boldsymbol{\Sigma}_i.$$

We note that $E(\boldsymbol{x}|\boldsymbol{x} \in \Pi_i) = \boldsymbol{\mu}_i$ and $\mathrm{Var}(\boldsymbol{x}|\boldsymbol{x} \in \Pi_i) = \boldsymbol{\Sigma}_i$ for $i = 1, \dots, k$. We denote the eigendecomposition of $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ having eigenvalues $\lambda_1 \geq \dots \geq \lambda_d \geq 0$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_d]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{x}_j - \boldsymbol{\mu} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}(z_{1j}, \dots, z_{dj})^{\mathrm{T}}$ for $j = 1, \dots, n$. Then, $(z_{1j}, \dots, z_{dj})^{\mathrm{T}}$ is a sphered data vector from a distribution with the identity covariance matrix; $E\{(z_{1j}, \dots, z_{dj})^{\mathrm{T}}\} = \boldsymbol{0}$ and $\mathrm{Var}\{(z_{1j}, \dots, z_{dj})^{\mathrm{T}}\} = \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ denotes the d-square identity matrix. The ith true PC score of $\boldsymbol{x}_j$ is given by

$$\boldsymbol{h}_i^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu}) = \lambda_i^{1/2} z_{ij} \text{ (hereafter called } s_{ij}).$$

We note that $\mathrm{Var}(s_{ij}) = \lambda_i$ for all i,j. Let $\Delta_i = ||\boldsymbol{\mu}_i||^2$ for $i = 1, \dots, k$, where $|| \cdot ||$ denotes the Euclidean norm. Here, we assume that

$$\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_k,$$

without loss of generality. We also assume that

$$\Delta_k = 0 \text{ (i.e., } \boldsymbol{\mu}_k = \boldsymbol{0}),$$

for the sake of simplicity.

Remark 1. When $\boldsymbol{\mu}_k \neq \boldsymbol{0}$, let $\boldsymbol{\mu}_i' = \boldsymbol{\mu}_i - \boldsymbol{\mu}_k$ for $i = 1, \dots, k$. Then, it holds that $\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \varepsilon_i \varepsilon_j (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \varepsilon_i \varepsilon_j (\boldsymbol{\mu}_i' - \boldsymbol{\mu}_j')(\boldsymbol{\mu}_i' - \boldsymbol{\mu}_j')^{\mathrm{T}}$. Hence, for any inference of $\boldsymbol{\Sigma}$ by the sample covariance matrix, one can assume $\boldsymbol{\mu}_k = \boldsymbol{0}$ without loss of generality.

As the sign of an eigenvector is arbitrary, we assume that $\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\mu}_i \geq 0$ for $i = 1, \dots, k-1$, without loss of generality. In addition, we assume the cluster means are more spread than the within class variation in the sense that:

Condition 1. $\frac{\max_{i=1,\dots,k} \lambda_{\max}(\boldsymbol{\Sigma}_i)}{\Delta_{k-1}} \to 0$ as $d \to \infty$.

Here, $\lambda_{\max}(\boldsymbol{M})$ denotes the largest eigenvalue of any positive-semidefinite matrix, $\boldsymbol{M}$. We consider clustering $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ into one of Π_is in HDLSS situations. When $k = 2$, Yata and Aoshima (2010) gave the following result: we denote the angle between two nonzero vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ by $\mathrm{Angle}(\boldsymbol{x},\boldsymbol{y}) = \cos^{-1}\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}/(||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||)\}$. By noting that $\boldsymbol{\mu}_2 = \boldsymbol{0}$, under Condition 1, it holds that as $d \to \infty$

$$\frac{\lambda_1}{\varepsilon_1 \varepsilon_2 \Delta_1} \to 1 \quad \text{and} \quad \mathrm{Angle}(\boldsymbol{h}_1, \boldsymbol{\mu}_1) \to 0, \tag{2}$$

from the fact that $\lambda_1/\Delta = \boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{h}_1/\Delta = \varepsilon_1\varepsilon_2(\boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\mu}_1)^2 + o(1)$ as $d \to \infty$ under Condition 1. Furthermore, for the normalized first PC score $s_{1j}/\lambda_1^{1/2}$ $(= z_{1j})$, it follows that

$$\operatorname*{plim}_{d\to\infty} \frac{s_{1j}}{\lambda_1^{1/2}} = \begin{cases} (\varepsilon_2/\varepsilon_1)^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_1, \\ -(\varepsilon_1/\varepsilon_2)^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_2, \end{cases} \tag{3}$$

for $j = 1, \dots, n$. Here, "plim" denotes the convergence in probability. This result is a special case of Theorem 2 in Section 3. See Remark 8. One would be able to cluster $\boldsymbol{x}_j$s into two groups if s_{1j} is accurately estimated in HDLSS situations.

In this article, we consider asymptotic properties of sample PC scores for Equation (1) in the HDLSS context that $d \to \infty$ while n is fixed. In Section 2, we first derive a geometric representation of HDLSS data taken from the two-class mixture model. With the help of the geometric representation, we give geometric consistency properties of sample PC scores in the HDLSS context. We show that PCA can cluster HDLSS data under certain conditions in a surprisingly explicit way. In Section 3, we investigate asymptotic behaviors of true PC scores for the $k(\geq 3)$-class mixture model and provide geometric consistency properties of sample PC scores when $k \geq 3$. In Section 4, we demonstrate the performance of clustering based on sample PC scores using gene expression datasets. We show that the real-HDLSS datasets hold the geometric consistency properties.

2 | PC SCORES FOR TWO-CLASS MIXTURE MODEL

In this section, we consider PC scores for the two-class ($k = 2$) mixture model.

2.1 | Preliminary

The sample covariance matrix is given by $\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}} = (n-1)^{-1}\sum_{j=1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_n)^{\mathrm{T}}$, where $\overline{\boldsymbol{x}}_n = n^{-1}\sum_{j=1}^{n}\boldsymbol{x}_j$ and $\overline{\boldsymbol{X}} = \overline{\boldsymbol{x}}_n\boldsymbol{1}_n^{\mathrm{T}}$ with $\boldsymbol{1}_n = (1, \dots, 1)^{\mathrm{T}} \in \mathbb{R}^n$. Then, we define the $n \times n$ dual sample covariance matrix by $\boldsymbol{S}_{\mathrm{D}} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}(\boldsymbol{X} - \overline{\boldsymbol{X}})$. We note that $\mathrm{rank}(\boldsymbol{S}_{\mathrm{D}}) \leq n - 1$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1} \geq 0$ be the eigenvalues of $\boldsymbol{S}_{\mathrm{D}}$. Then, we define the eigendecomposition of $\boldsymbol{S}_{\mathrm{D}}$ by

$$\boldsymbol{S}_{\mathrm{D}} = \sum_{i=1}^{n-1} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^{\mathrm{T}},$$

where $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \dots, \hat{u}_{in})^{\mathrm{T}}$ denotes a unit eigenvector corresponding to $\hat{\lambda}_i$. As the sign of $\hat{\boldsymbol{u}}_i$s is arbitrary, we assume $\hat{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{z}_i \geq 0$ for all i without loss of generality, where $\boldsymbol{z}_i$ is defined by $\boldsymbol{z}_i = (z_{i1}, \dots, z_{in})^{\mathrm{T}}$. Note that $\boldsymbol{S}$ and $\boldsymbol{S}_{\mathrm{D}}$ share the nonzero eigenvalues. Let

$$\hat{z}_{ij} = \hat{u}_{ij}n^{1/2} \quad \text{for } i = 1, \dots, n-1;\ j = 1, \dots, n.$$

We note that $\hat{z}_{ij}$ is an estimate of $s_{ij}/\lambda_i^{1/2}$ $(=z_{ij})$ for $i = 1, \dots, n-1; j = 1, \dots, n$ from the facts that

$$\hat{z}_{ij} = \{n/(n-1)\}^{1/2}\hat{\boldsymbol{h}}_i^{\mathrm{T}}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_n)/\hat{\lambda}_i^{1/2} \quad \text{and} \quad \sum_{j=1}^{n}\hat{z}_{ij}^2/n = 1 \quad \text{if } \hat{\lambda}_i > 0,$$

where $\hat{\boldsymbol{h}}_i$ denotes a unit eigenvector of $\boldsymbol{S}$ corresponding to $\hat{\lambda}_i$. Let $\boldsymbol{X}_0 = \boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{1}_n^{\mathrm{T}}$ and $\boldsymbol{P}_n = \boldsymbol{I}_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^{\mathrm{T}}$. We note that $\boldsymbol{S}_{\mathrm{D}} = \boldsymbol{P}_n\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{X}_0\boldsymbol{P}_n/(n-1)$. We consider the sphericity condition: $\mathrm{tr}(\boldsymbol{\Sigma}^2)/\mathrm{tr}(\boldsymbol{\Sigma})^2 \to 0$ as $d \to \infty$. Note that the sphericity condition is equivalent to "$\lambda_1/\mathrm{tr}(\boldsymbol{\Sigma}) \to 0$ as $d \to \infty$." When one can assume that $\boldsymbol{X}$ is Gaussian or $\boldsymbol{Z} = (z_{ij})$ is ρ-mixing and the fourth moments of each variable in $\boldsymbol{Z}$ are uniformly bounded, under the sphericity condition, Jung and Marron (2009) suggested a geometric representation as follows:

$$\operatorname*{plim}_{d\to\infty}\frac{\boldsymbol{X}_0^{\mathrm{T}}\boldsymbol{X}_0}{\mathrm{tr}(\boldsymbol{\Sigma})}=\boldsymbol{I}_n,\quad \text{so that } \operatorname*{plim}_{d\to\infty}\frac{(n-1)\boldsymbol{S}_{\mathrm{D}}}{\mathrm{tr}(\boldsymbol{\Sigma})}=\boldsymbol{P}_n. \tag{4}$$

Remark 2. Yata and Aoshima (2012) showed that Equation (4) holds under the sphericity condition and $\mathrm{Var}(||\boldsymbol{x}_j-\boldsymbol{\mu}||^2)/\mathrm{tr}(\boldsymbol{\Sigma})^2\to 0$ as $d\to\infty$.

From Equation (4), we observe that the eigenvalue becomes deterministic as the dimension increases, whereas the eigenvector of $\boldsymbol{S}_{\mathrm{D}}$ does not uniquely determine the direction. In addition, Hellton and Thoresen (2017) present asymptotic properties of the sample PC scores when $\boldsymbol{Z}$ is ρ-mixing. We note that Equation (1) does not presuppose the assumption that $\boldsymbol{X}$ is Gaussian or $\boldsymbol{Z}$ is ρ-mixing. See section 4.1.1 in Qiao, Zhang, Liu, Todd, and Marron (2010) for details. In the present article, we present new asymptotic properties of the sample PC for Equation (1).

2.2 | Geometric representation and consistency property of PC scores when $k=2$

We will find a geometric representation for Equation (1) and the finding is completely different from Equation (4). We assume the following conditions:

Condition 2. $\frac{\max_{i=1,\ldots,k}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta_{k-1}^2}\to 0$ as $d\to\infty$.

Condition 3. $\frac{\max_{i=1,\ldots,k}\mathrm{Var}(||\boldsymbol{x}-\boldsymbol{\mu}_i||^2|\boldsymbol{x}\in\Pi_i)}{\Delta_{k-1}^2}\to 0$ as $d\to\infty$.

Condition 4. $\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)-\mathrm{tr}(\boldsymbol{\Sigma}_j)}{\Delta_{k-1}}\to 0$ as $d\to\infty$ for all $i,j=1,\ldots,k(i<j)$.

Remark 3. Condition 2 is stronger than Condition 1 as it holds that $\{\lambda_{\max}(\boldsymbol{\Sigma}_i)\}^2\le\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ for $i=1,\ldots,k$. Let $\beta(>0)$ be a constant such that $\liminf_{d\to\infty}(\Delta_{k-1}/d^{\beta})>0$. Let $\lambda_{i1}\ge\cdots\ge\lambda_{id}\ge 0$ be eigenvalues of $\boldsymbol{\Sigma}_i$ for $i=1,\ldots,k$. For a spiked model such as

$$\lambda_{ij}=a_{ij}d^{\alpha_{ij}}\ (j=1,\ldots,t_i)\quad\text{and}\quad\lambda_{ij}=c_{ij}\ (j=t_i+1,\ldots,d),$$

with positive constants, a_{ij}, c_{ij}, and α_{ij} (not depending on d), and a positive integer t_i (not depending on d), Condition 1 holds when $\alpha_{i1}<\beta$ for $i=1,\ldots,k$. Also, Condition 2 holds when $\beta>1/2$ and $\alpha_{i1}<\beta$ for $i=1,\ldots,k$. See Yata and Aoshima (2012) for the details of the spiked model.

Remark 4. If Π_is are Gaussian, it holds that $\mathrm{Var}(||\boldsymbol{x}-\boldsymbol{\mu}_i||^2|\boldsymbol{x}\in\Pi_i)=O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$ for $i=1,\ldots,k$, so that Condition 3 naturally holds under Condition 2.

Remark 5. When $k=2$, Condition 4 holds if $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2)\to 1$ as $d\to\infty$ and $\liminf_{d\to\infty}\{\Delta_1/\mathrm{tr}(\boldsymbol{\Sigma})\}>0$.

We define that

$$r_j=(-1)^{i+1}(1-\eta_i)\text{ according to }\boldsymbol{x}_j\in\Pi_i\quad\text{for } j=1,\ldots,n.$$

The following result gives a geometric representation for Equation (1) when $k=2$.

Theorem 1. *Assume $\Delta_1/tr(\boldsymbol{\Sigma}) \to c(>0)$ as $d \to \infty$. Under Conditions 2–4, it holds*

$$\underset{d\to\infty}{\text{plim}} \frac{(n-1)\boldsymbol{S}_{\text{D}}}{tr(\boldsymbol{\Sigma})} = c\boldsymbol{r}\boldsymbol{r}^{\text{T}} + (1-\varepsilon_1\varepsilon_2 c)\boldsymbol{P}_n, \tag{5}$$

where $\boldsymbol{r} = (r_1, \dots, r_n)^{\text{T}}$.

When $\boldsymbol{S}_{\text{D}} \neq \boldsymbol{O}$, we note that $\hat{\boldsymbol{u}}_1^{\text{T}}\boldsymbol{1}_n = 0$, so that $\hat{\boldsymbol{u}}_1^{\text{T}}\boldsymbol{P}_n = \hat{\boldsymbol{u}}_1^{\text{T}}$. Thus from Equation (5), the first eigenvector of $\boldsymbol{S}_{\text{D}}$ uniquely determines the direction. In fact, by noting $\boldsymbol{r}^{\text{T}}\boldsymbol{1}_n = 0$ and $||\boldsymbol{r}||^2 = n\eta_1\eta_2$, we have the following results for the first eigenvector and PC scores when $k = 2$. Using Corollary 1, one can cluster $\boldsymbol{x}_j$s into two groups by the sign of $\hat{z}_{1j}$s:

Corollary 1. *Under Conditions 2–4, it holds that for $n_i > 0, i = 1, 2$*

$$\underset{d\to\infty}{\text{plim}}\ \hat{\boldsymbol{u}}_1 = \frac{\boldsymbol{r}}{(n\eta_1\eta_2)^{1/2}} \quad \text{and}$$

$$\underset{d\to\infty}{\text{plim}}\ \hat{z}_{1j} = \begin{cases} (\eta_2/\eta_1)^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_1, \\ -(\eta_1/\eta_2)^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_2, \end{cases} \quad \text{for } j = 1, \dots, n.$$

We considered an easy example such as $\Pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2,$ with $\boldsymbol{\mu}_1 = \boldsymbol{1}_d$, $\boldsymbol{\mu}_2 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = (0.3^{|i-j|^{1/3}})$, and $\boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B} = \text{diag}[-\{0.5 + 1/(d+1)\}^{1/2}, \{0.5 + 2/(d+1)\}^{1/2}, \dots, (-1)^d\{0.5 + d/(d+1)\}^{1/2}]$. We note that $\Delta_1 = d$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ but $\text{tr}(\boldsymbol{\Sigma}_1) = \text{tr}(\boldsymbol{\Sigma}_2) = d$. Then, Conditions 2–4 hold. We set $n_1 = 1$ and $n_2 = 2$. We took $n = 3$ samples as $\boldsymbol{x}_1 \in \Pi_1$ and $\boldsymbol{x}_2, \boldsymbol{x}_3 \in \Pi_2$. In Figure 1, we displayed scatter plots of 20 independent pairs of $\pm\hat{\boldsymbol{u}}_1$ when (a) $d = 5$, (b) $d = 50$, (c) $d = 500$, and (d) $d = 5{,}000$. We denoted $\boldsymbol{r} = (2/3, -1/3, -1/3)^{\text{T}}$ by the solid line and $\boldsymbol{1}_n = (1, 1, 1)^{\text{T}}$ by the dotted line. We note that $\text{Angle}(\hat{\boldsymbol{u}}_1, \boldsymbol{1}_n) = \pi/2$ when $\boldsymbol{S}_{\text{D}} \neq \boldsymbol{O}$. We observed that all the plots of $\pm\hat{\boldsymbol{u}}_1$ gather on the surface of the orthogonal complement of $\boldsymbol{1}_n$. Also, the plots appeared close to $\boldsymbol{r}$ as d increases. Thus, one can cluster $\boldsymbol{x}_j$s into two groups by the sign of $\hat{z}_{1j}$s.

Next, we investigated robustness of Corollary 1 against Condition 4 by some simulation studies. Let $\Delta_\Sigma = |\text{tr}(\boldsymbol{\Sigma}_1) - \text{tr}(\boldsymbol{\Sigma}_2)|$. We considered an easy example such as $\Pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2,$ with $\boldsymbol{\mu}_1 = (1, \dots, 1, 0, \dots, 0)^{\text{T}}$ whose first $\lceil d^{3/4} \rceil$ elements are 1, $\boldsymbol{\mu}_2 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = \gamma(0.3^{|i-j|^{1/3}})$, and $\boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, where $\gamma \geq 1$. Here, $\lceil \cdot \rceil$ denotes the ceiling function. Note that $\Delta_\Sigma = (\gamma - 1)d$. We set $d = 5{,}000$, $n = 10$, $n_1 = 4$, and $n_2 = 6$. We took $n = 10$ samples as $\boldsymbol{x}_1, \dots, \boldsymbol{x}_4 \in \Pi_1$ and $\boldsymbol{x}_5, \dots, \boldsymbol{x}_{10} \in \Pi_2$. In Figure 2, we displayed scatter plots of $(\hat{z}_{1j}, \hat{z}_{2j})$, $j = 1, \dots, n$, when (a) $\gamma = 1 + 2d^{-1/2}$, (b) $\gamma = 1 + 2d^{-1/4}$, and (c) $\gamma = 3$. From Corollary 1

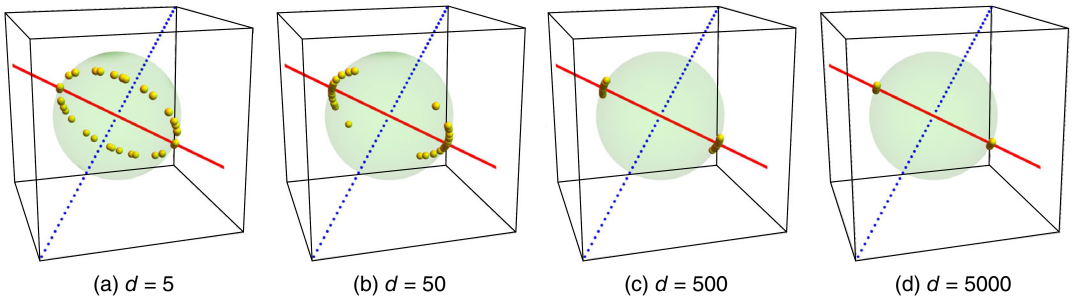
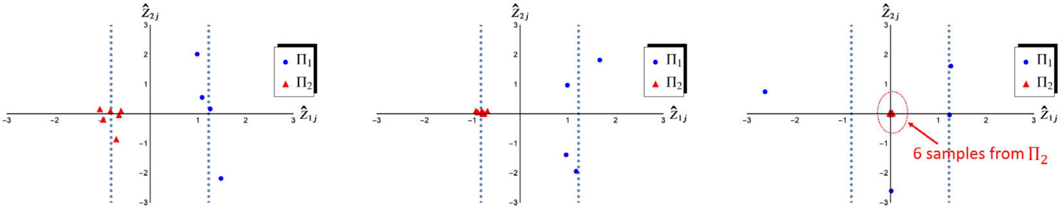

(a) $d = 5$ (b) $d = 50$ (c) $d = 500$ (d) $d = 5000$

FIGURE 1 Toy example to illustrate the geometric representation of $\pm\hat{\boldsymbol{u}}_1$ on the unit sphere when $k = 2$ and $n = 3$. We plotted 20 independent pairs of $\pm\hat{\boldsymbol{u}}_1$ when $\boldsymbol{x}_1 \in \Pi_1$ and $\boldsymbol{x}_2, \boldsymbol{x}_3 \in \Pi_2$. The solid line denotes $\boldsymbol{r} = (2/3, -1/3, -1/3)^{\text{T}}$ and the dotted line denotes $\boldsymbol{1}_n = (1, 1, 1)^{\text{T}}$ [Colour figure can be viewed at wileyonlinelibrary.com]



(a) $\Delta_1/\Delta_\Sigma \approx d^{1/4}/2$

(b) $\Delta_1/\Delta_\Sigma \approx 1/2$

(c) $\Delta_1/\Delta_\Sigma \approx d^{-1/4}/2$

FIGURE 2 Toy example to illustrate asymptotic behaviors of the estimated principal component scores when $k = 2$. We plotted $(\hat{z}_{1j}, \hat{z}_{2j})$ which is denoted by small circles when $\boldsymbol{x}_j \in \Pi_1$ and by small triangles when $\boldsymbol{x}_j \in \Pi_2$. The theoretical convergent points, $(3/2)^{1/2}$ and $-(3/2)^{1/2}$, are denoted by dotted lines [Colour figure can be viewed at wileyonlinelibrary.com]

we denoted $(3/2)^{1/2}$ and $-(2/3)^{1/2}$ by dotted lines. Note that $\Delta_\Sigma/\Delta_1 \approx 2d^{-1/4}$ for (a), $\Delta_\Sigma/\Delta_1 \approx 2$ for (b), and $\Delta_\Sigma/\Delta_1 \approx 2d^{1/4}$ for (c). Thus, Condition 4 holds for (a), while it does not hold for (b) and (c). For (a) and (b), we observed that the estimated PC scores give good performances. On the other hand, the first PC score did not gather around $(3/2)^{1/2}$ or $-(2/3)^{1/2}$ for (c). However, $(\hat{z}_{1j}, \hat{z}_{2j})$s were concentrated on the origin $(0, 0)$ for $\boldsymbol{x}_j \in \Pi_2$.

When $\Delta_1/\Delta_\Sigma \to 0$ as $d \to \infty$, we give the following result to explain the reason of the phenomenon in Figure 2c. Under the assumptions of Proposition 1, one can cluster $\boldsymbol{x}_j$s into two groups by the size of $\hat{z}_{ij}$s even when Condition 4 is not met:

Proposition 1. *Assume $k = 2$, $n_{l_*} \geq 2$ and $n_{l'_*} \geq 1$, where $l_*(\neq l'_*)$ is an integer such that $tr(\boldsymbol{\Sigma}_{l_*}) >$* $\mathrm{tr}(\boldsymbol{\Sigma}_{l'_*})$*. Assume also that* $\max_{l=1,2} tr(\boldsymbol{\Sigma}_l^2)/\Delta_\Sigma^2 \to 0$, $\max_{l=1,2}\mathrm{Var}(||\boldsymbol{x} - \boldsymbol{\mu}_l||^2|\boldsymbol{x} \in \Pi_l)/\Delta_\Sigma^2 \to 0$ *and* $\Delta_1/\Delta_\Sigma \to 0$ *as $d \to \infty$. Then, it holds that*

$$\operatorname*{plim}_{d\to\infty} |\hat{z}_{ij}| > 0 \ \textit{when} \ \ \boldsymbol{x}_j \in \Pi_{l_*} \ \textit{for some} \ i \in [1, n_{l_*} - 1] \quad \textit{and}$$

$$\operatorname*{plim}_{d\to\infty} \hat{z}_{ij} = 0 \ \textit{when} \ \ \boldsymbol{x}_j \in \Pi_{l'_*} \ \textit{for} \ i = 1, \dots, n_{l_*} - 1.$$

Remark 6. For $k \geq 3$, we do not give any consistency property when Condition 4 is not met because the sufficient conditions of Proposition 1 become quite complicated for $k \geq 3$. Detailed study for the case when $k \geq 3$ is left for a future work.

The assumptions of Proposition 1 hold for (c) of Figure 2. Thus, $(\hat{z}_{1j}, \hat{z}_{2j})$s were concentrated on the origin $(0, 0)$ for $\boldsymbol{x}_j \in \Pi_2$ in (c).

3 | PC SCORES FOR MULTICLASS MIXTURE MODEL

In this section, we consider PC scores for the $k(\geq 3)$-class mixture model.

3.1 | Asymptotic behaviors of true PC scores when $k \geq 3$

Let

$$\varepsilon_{(0)} = 0 \quad \text{and} \quad \varepsilon_{(i)} = \sum_{j=1}^{i} \varepsilon_j \quad \text{for } i = 1, \dots, k.$$

We assume the following condition:

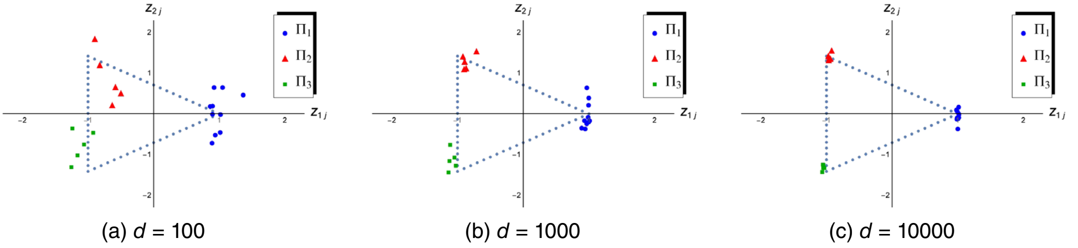
Condition 5. Angle$(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \to \frac{\pi}{2}$ and $\frac{\Delta_j}{\Delta_i} \to 0$ as $d \to \infty$ for $i, j = 1, \dots, k-1 (i < j)$.

Remark 7. We consider the case when all elements of $\boldsymbol{\mu}_i$s are constants (not depending on d) such as $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ip}, 0, \dots, 0)$ with $\mu_{is} \neq 0$ (not depending on d) for $s = 1, \dots, p$. If all elements of $\boldsymbol{\mu}_i$s are constants, the condition "Angle$(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \to \pi/2$ as $d \to \infty$" holds for $i < j$ under $\Delta_j/\Delta_i \to 0$ as $d \to \infty$, so that Condition 5 holds under $\Delta_{i+1}/\Delta_i \to 0$ as $d \to \infty$ for $i = 1, \dots, k-2$. See the settings of Figures 3 and 4. Note that $\Delta_1 \gg \cdots \gg \Delta_{k-1}$ under Condition 5. We emphasize that Conditions 1–4 become strict as k increases under Condition 5.

We have the following results.

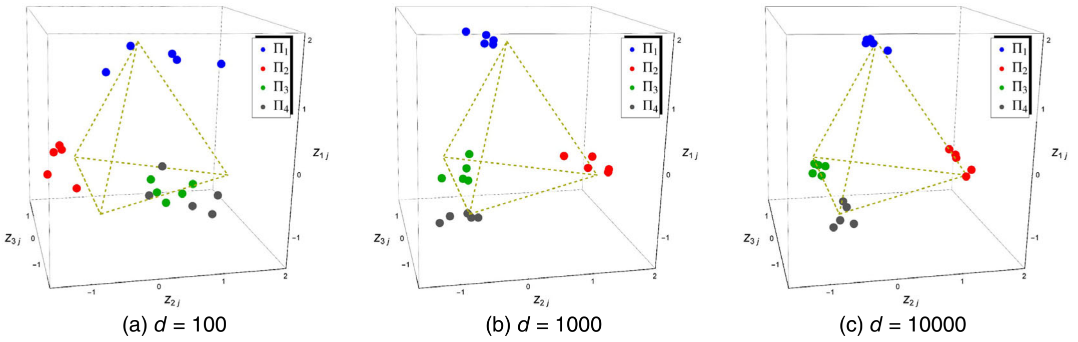
Theorem 2. *Under Conditions 1 and 5, it holds that for $i = 1, \dots, k-1; j = 1, \dots, n$*

$$\operatorname*{plim}_{d\to\infty} \frac{s_{ij}}{\lambda_i^{1/2}} = \begin{cases} 0 & \textit{when } i \geq 2 \textit{ and } \boldsymbol{x}_j \in \cup_{m=1}^{i-1} \Pi_m, \\ \left(\frac{1-\varepsilon_{(i)}}{\varepsilon_i (1-\varepsilon_{(i-1)})} \right)^{1/2} & \textit{when } \boldsymbol{x}_j \in \Pi_i, \\ -\left(\frac{\varepsilon_i}{(1-\varepsilon_{(i)})(1-\varepsilon_{(i-1)})} \right)^{1/2} & \textit{when } \boldsymbol{x}_j \in \cup_{m=i+1}^{k} \Pi_m. \end{cases} \qquad (6)$$



(a) d = 100 (b) d = 1000 (c) d = 10000

FIGURE 3 Toy example to illustrate the asymptotic behaviors of true principal component scores when $k = 3$. We plotted (z_{1j}, z_{2j}) which is denoted by small circles when $\boldsymbol{x}_j \in \Pi_1$, by small triangles when $\boldsymbol{x}_j \in \Pi_2$, and by small squares when $\boldsymbol{x}_j \in \Pi_3$. The dashed triangle consists of three vertices, namely, $(1, 0)$, $(-1, 2^{1/2})$, and $(-1, -2^{1/2})$, which are theoretical convergent points [Colour figure can be viewed at wileyonlinelibrary.com]



(a) d = 100 (b) d = 1000 (c) d = 10000

FIGURE 4 Toy example to illustrate the asymptotic behaviors of true principal component scores when $k = 4$. We plotted (z_{1j}, z_{2j}, z_{3j}). The dashed triangular pyramid was given by Equation (6) with $k = 4$ [Colour figure can be viewed at wileyonlinelibrary.com]

Remark 8. The consistency in Equation (3) is equivalent to Equation (6) with $k = 2$ and $i = 1$.

Corollary 2. *Under Conditions* 1 *and* 5*, it holds that for* $i = 1, \dots, k-1$

$$\frac{\lambda_i}{\varepsilon_i(1-\varepsilon_{(i)})\Delta_i/(1-\varepsilon_{(i-1)})} \to 1 \quad \textit{and} \quad \mathrm{Angle}(\boldsymbol{h}_i, \boldsymbol{\mu}_i) \to 0 \quad \textit{as } d \to \infty.$$

For example, when $k = 3$, from Equation (6), we have that for $j = 1, \dots, n$

$$\operatorname*{plim}_{d\to\infty} \frac{s_{1j}}{\lambda_1^{1/2}} = \begin{cases} \{(1-\varepsilon_1)/\varepsilon_1\}^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_1, \\ -\{\varepsilon_1/(1-\varepsilon_1)\}^{1/2} & \text{when } \boldsymbol{x}_j \notin \Pi_1 \end{cases} \quad \text{and}$$

$$\operatorname*{plim}_{d\to\infty} \frac{s_{2j}}{\lambda_2^{1/2}} = \begin{cases} 0 & \text{when } \boldsymbol{x}_j \in \Pi_1, \\ [\varepsilon_3/\{\varepsilon_2(1-\varepsilon_1)\}]^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_2, \\ -[\varepsilon_2/\{\varepsilon_3(1-\varepsilon_1)\}]^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_3. \end{cases}$$

One can check whether $\boldsymbol{x}_j \in \Pi_1$ or not by the first PC score. If $\boldsymbol{x}_j \notin \Pi_1$, one can check whether $\boldsymbol{x}_j \in \Pi_2$ or $\boldsymbol{x}_j \in \Pi_3$ by the second PC score. In general, one can cluster $\boldsymbol{x}_j$s using at most the first $k-1$ PC scores.

We considered a toy example such as $\Pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, \dots, 4$, where $\boldsymbol{\mu}_1 = \boldsymbol{1}_d$, $\boldsymbol{\mu}_2 = (1, \dots, 1, 0, \dots, 0)^{\mathrm{T}}$ whose first $\lceil d^{3/4} \rceil$ elements are 1, $\boldsymbol{\mu}_3 = (1, \dots, 1, 0, \dots, 0)^{\mathrm{T}}$ whose first $\lceil d^{1/2} \rceil$ elements are 1, and $\boldsymbol{\mu}_4 = \boldsymbol{0}$. We set $\boldsymbol{\Sigma}_1 = (0.3^{|i-j|^{1/3}})$, $\boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, $\boldsymbol{\Sigma}_3 = 0.8\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_4 = 1.2\boldsymbol{\Sigma}_2$, where $\boldsymbol{B}$ is defined in Section 2.2. Then, Conditions 1 and 5 hold. We first considered the case when $k = 3 : \Pi_i, i = 1, 2, 3$, having $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (1/2, 1/4, 1/4)$. We set $n = 20$ and $(n_1, n_2, n_3) = (10, 5, 5)$. From Theorem 2, one can expect that (z_{1j}, z_{2j}) $(= (s_{1j}/\lambda_1^{1/2}, s_{2j}/\lambda_2^{1/2}))$ becomes close to $(1, 0)$ when $\boldsymbol{x}_j \in \Pi_1$, $(-1, 2^{1/2})$ when $\boldsymbol{x}_j \in \Pi_2$, and $(-1, -2^{1/2})$ when $\boldsymbol{x}_j \in \Pi_3$. In Figure 3, we displayed scatter plots of (z_{1j}, z_{2j}), $j = 1, \dots, n$, when (a) $d = 100$, (b) $d = 1{,}000$, and (c) $d = 10{,}000$. We observed that the scatter plots appear close to those three vertices as d increases.

Next, we considered the case when $k = 4 : \Pi_i, i = 1, \dots, 4$, having $\varepsilon_1 = \cdots = \varepsilon_4 = 1/4$. We set $n = 20$ and $n_1 = \cdots = n_4 = 5$. In Figure 4, we displayed scatter plots of (z_{1j}, z_{2j}, z_{3j}), $j = 1, \dots, n$, when (a) $d = 100$, (b) $d = 1{,}000$ and (c) $d = 10{,}000$. From Theorem 2, we displayed the triangular pyramid given by Equation (6) with $k = 4$. As expected theoretically, we observed that the scatter plots appear close to four vertices of the triangular pyramid as d increases. They seemed to converge slower in Figure 4 than in Figure 3. This is because the conditions of Theorem 2 become strict as k increases. See Remark 7.

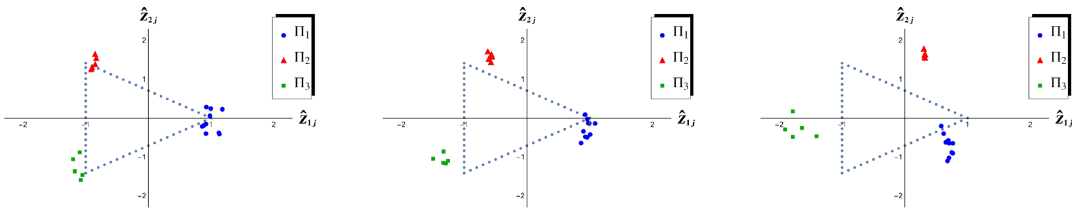
3.2 | Consistency property of PC scores when $k \geq 3$

Let

$$\eta_{(0)} = 0 \quad \text{and} \quad \eta_{(i)} = \sum_{j=1}^{i} \eta_j \text{ for } i = 1, \dots, k.$$

We assume the following condition:

Condition 6. $\frac{\max_{i=1,\dots,k-2}(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\mu}_i)}{\Delta_{k-1}^2} \to 0$ as $d \to \infty$ for $j = 1, \dots, k$.



(a) Angle $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 0.352\pi$ (b) Angle $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 0.282\pi$ (c) Angle $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 0.148\pi$

FIGURE 5 Toy example to illustrate asymptotic behaviors of the estimated principal component scores when $k = 3$. We plotted $(\hat{z}_{1j}, \hat{z}_{2j})$ which is denoted by small circles when $\boldsymbol{x}_j \in \Pi_1$, by small triangles when $\boldsymbol{x}_j \in \Pi_2$, and by small squares when $\boldsymbol{x}_j \in \Pi_3$. The dashed triangle consists of three vertices, namely, $(1, 0)$, $(-1, 2^{1/2})$, and $(-1, -2^{1/2})$, which are the theoretical convergent points [Colour figure can be viewed at wileyonlinelibrary.com]

Remark 9. From the fact that $\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\Sigma}_j \boldsymbol{\mu}_i \le \Delta_i \lambda_{\max}(\boldsymbol{\Sigma}_j)$, Condition 6 holds under $\Delta_1 \lambda_{\max}(\boldsymbol{\Sigma}_j)/\Delta_{k-1}^2 \to 0$ as $d \to \infty$ for $j = 1, \dots, k$.

As for the estimated PC scores, we have the following result. From Theorem 3, one can cluster $\boldsymbol{x}_j$s into k groups by the elements of $\hat{\boldsymbol{u}}_i$, $i = 1, \dots, k-1$:

Theorem 3. *Under Conditions 2–6, it holds that for $n_l > 0$, $l = 1, \dots, k$*

$$\operatorname*{plim}_{d\to\infty} \hat{z}_{ij} = \begin{cases} 0 & \text{when } i \ge 2 \text{ and } \boldsymbol{x}_j \in \bigcup_{m=1}^{i-1} \Pi_m, \\ \left(\dfrac{1-\eta_{(i)}}{\eta_i(1-\eta_{(i-1)})} \right)^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_i, \\ -\left(\dfrac{\eta_i}{(1-\eta_{(i)})(1-\eta_{(i-1)})} \right)^{1/2} & \text{when } \boldsymbol{x}_j \in \bigcup_{m=i+1}^{k} \Pi_m, \end{cases} \tag{7}$$

for $i = 1, \dots, k-1; j = 1, \dots, n$.

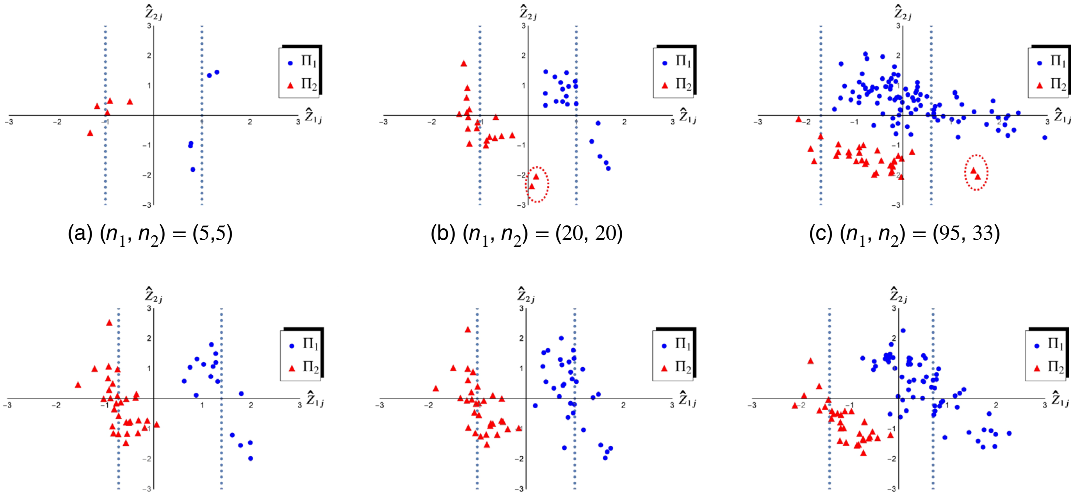
Condition 5 is essential for the consistency properties given in Theorems 2 and 3. We investigated the robustness of Theorem 3 against Condition 5 by some simulation studies. We considered a toy example such as $\Pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, 2, 3$, where $\boldsymbol{\mu}_1 = \boldsymbol{1}_d$, $\boldsymbol{\mu}_2 = (1, \dots, 1, 0, \dots, 0)^{\mathrm{T}}$ whose first $\lceil d/\zeta \rceil$ elements are 1, $\boldsymbol{\mu}_3 = \boldsymbol{0}, \boldsymbol{\Sigma}_1 = (0.3^{|i-j|^{1/3}}), \boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, and $\boldsymbol{\Sigma}_3 = (0.4^{|i-j|^{1/3}})$. We set $d = 5{,}000$, $n = 20$, and $(n_1, n_2, n_3) = (10, 5, 5)$. In Figure 5, we displayed scatter plots of $(\hat{z}_{1j}, \hat{z}_{2j})$, $j = 1, \dots, n$, when (a) $\zeta = 1/5$, (b) $\zeta = 2/5$, and (c) $\zeta = 4/5$. Also, we displayed the triangle given by Equation (7) with $k = 3$. Note that Angle$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 0.352\pi$ and $\Delta_2/\Delta_1 = 1/5$ for (a), Angle$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 0.282\pi$ and $\Delta_2/\Delta_1 = 2/5$ for (b), and Angle$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 0.148\pi$ and $\Delta_2/\Delta_1 = 4/5$ for (c). For (a) and (b), we observed that the estimated PC scores give good performances. On the other hand, the estimated PC scores seemed not to converge to the theoretical points for (c). This is because Condition 5 is not met. However, we could find three separate clusters for Π_i, $i = 1, 2, 3$. See Appendix B for the reason.

4 | REAL-DATA EXAMPLES

We demonstrate the performance of clustering, based on sample PC scores, using gene expression datasets.

4.1 | Clustering when $k = 2$

We analyzed microarray data by Chiaretti et al. (2004) in which the dataset consists of 12,625 (=d) genes and 128 samples. The dataset has two tumor cellular subtypes, Π_1 : B cell (95 samples) and Π_2 : T cell (33 samples). Refer to Jeffery, Higgins, and Culhane (2006) as well. We checked behaviors of the PC scores using several samples from the two tumor cellular subtypes. We considered three cases: (a) $n = 10$ samples consist of the first five samples from both Π_1 and Π_2 (i.e., $n_1 = 5$ and $n_2 = 5$), (b) $n = 40$ samples consist of the first 20 samples from both Π_1 and Π_2 (i.e., $n_1 = 20$ and $n_2 = 20$), and (c) $n = 128$ samples consist of $n_1 = 95$ samples from Π_1 and $n_2 = 33$ samples from Π_2. In the top panels of Figure 6, we displayed scatter plots of the first two PC scores, $(\hat{z}_{1j}, \hat{z}_{2j})$s, for (a), (b), and (c). From Corollary 1, we denoted $(\eta_2/\eta_1)^{1/2}$ and $-(\eta_1/\eta_2)^{1/2}$ by dotted lines. For (a), we observed that the estimated PC scores give good performances. The first PC scores gathered around $(\eta_2/\eta_1)^{1/2}$ or $-(\eta_1/\eta_2)^{1/2}$. For (b), the estimated PC scores gave adequate performances except for the two points from Π_2. Those two samples, which are the ninth and twentieth samples of Π_2, are probably outliers. In fact, the two points are far from the cluster of Π_2. The other 38 samples were perfectly classified into the two groups by the sign of the first PC scores. As for (c), although there seemed to be two clusters except for the two samples, we could not classify the dataset by the sign of the first PC scores. This is probably because η_1 and η_2 are unbalanced. From Equation (2), when the mixing proportions are unbalanced, λ_1 becomes small. The first eigenspace was possibly affected by the other eigenspaces, so that the first PC scores appear in the wrong direction. We tested the clustering except for the outlying two samples. We used the remaining 31 samples for Π_2. We considered the following three cases for samples from Π_1: (d) the first 16 samples from Π_1, so that $n_1 = 16, n_2 = 31, n = 47$, and $\eta_1/\eta_2 \approx 0.5$; (e) the first 31 samples from Π_1, so that $n_1 = 31, n_2 = 31, n = 62$, and $\eta_1/\eta_2 = 1$; and (f) the first 62 samples from Π_1, so that $n_1 = 62, n_2 = 31, n = 93$, and $\eta_1/\eta_2 = 2$. In the bottom panels of Figure 6, we displayed scatter



(a) $(n_1, n_2) = (5,5)$

(b) $(n_1, n_2) = (20, 20)$

(c) $(n_1, n_2) = (95, 33)$

(d) $(n_1, n_2) = (16, 31)$

(e) $(n_1, n_2) = (31, 31)$

(f) $(n_1, n_2) = (62, 31)$

FIGURE 6 Scatter plots of the first two principal component scores, supposing $k = 2$ in the dataset of Chiaretti et al. (2004). We denoted them by small circles when $\boldsymbol{x}_j \in \Pi_1$ and by small triangles when $\boldsymbol{x}_j \in \Pi_2$. The theoretical convergent points, namely, $(\eta_2/\eta_1)^{1/2}$ and $-(\eta_1/\eta_2)^{1/2}$, are denoted by dotted lines. The two samples, encircled by dots in (b) and (c), are probably outliers [Colour figure can be viewed at wileyonlinelibrary.com]

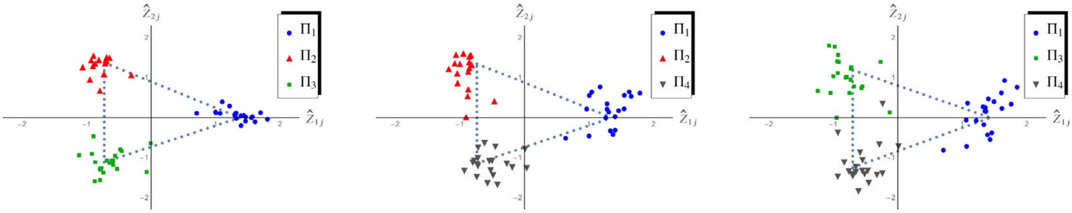
plots of $(\hat{z}_{1j}, \hat{z}_{2j})$s for (d), (e), and (f). For (d) and (e), we observed that the estimated PC scores give good performances. As for (f), although there seemed to be two clusters, we could not classify the dataset by the sign of the first PC scores. Note that η_1 and η_2 are unbalanced in (d) and (f). Even though (d) is an unbalanced case, the estimated PC scores worked well for the case. We had an estimate for the ratio of the largest eigenvalues, $\lambda_{\max}(\boldsymbol{\Sigma}_1)/\lambda_{\max}(\boldsymbol{\Sigma}_2)$, as 1.598 by the noise-reduction methodology given by Yata and Aoshima (2012). The first eigenspace of $\boldsymbol{\Sigma}$ in (d) is less affected by the first eigenspace of $\boldsymbol{\Sigma}_i$s than in (f) as $\boldsymbol{\Sigma} = \varepsilon_1\varepsilon_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} + \varepsilon_1\boldsymbol{\Sigma}_1 + \varepsilon_2\boldsymbol{\Sigma}_2$. This is probably the reason why the estimated PC scores gave good performances even in (d).

4.2 | Clustering when $k \geq 3$

We analyzed microarray data by Bhattacharjee et al. (2001) in which the dataset consisted of five lung carcinomas types with $d = 3{,}312$. We only used four classes as Π_1 : pulmonary carcinoids (20 samples), Π_2 : normal lung (17 samples), Π_3 : squamous cell lung carcinomas (21 samples), and Π_4 : adenocarcinomas (20 samples), so that $n_1 = 20, n_2 = 17, n_3 = 21$, and $n_4 = 20$. Note that Π_4 originally had 139 samples. We used only the first 20 samples from Π_4 in order to keep balance in sample sizes with the other classes. We first considered clustering when $k = 3$ under the following setups: (a) the dataset consists of Π_1, Π_2, and Π_3 ($n = 58$); (b) the dataset consists of Π_1, Π_2, and Π_4 ($n = 57$); and (c) the dataset consists of Π_1, Π_3, and Π_4 ($n = 61$). In Figure 7, we displayed scatter plots of the first two PC scores, $(\hat{z}_{1j}, \hat{z}_{2j})$s, for each of (a), (b), and (c). Also, we displayed the triangle given by Equation (7) with $k = 3$ using Theorem 3. We observed that the estimated PC scores give good performances. The three clusters gathered around each vertex for (a), (b), and (c).

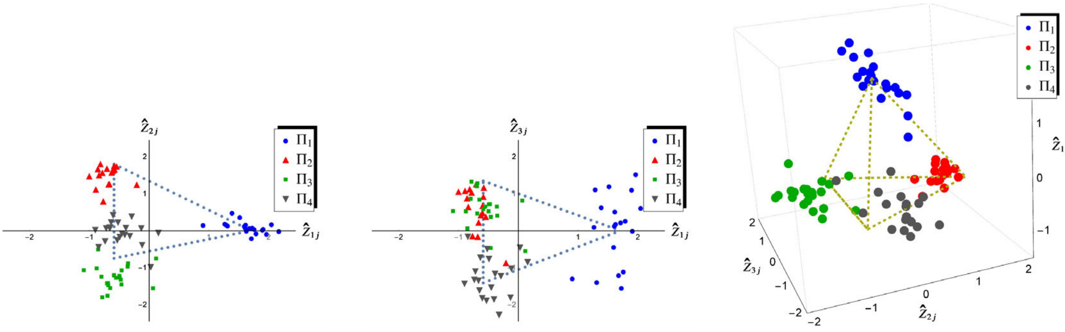
Next, we considered clustering when $k = 4$: $\Pi_i, i = 1, \dots, 4$, so that $n = 78$. In Figure 8, we displayed scatter plots of the first three PC scores. The dataset seemed not to converge to the theoretical convergent points given by Equation (7) in Theorem 3. This is probably because the conditions of Theorem 3 become strict as k increases. See Remark 7. Thus, the convergence is slower than in the case when $k = 3$ as in Figure 7. However, there seemed to be four separate clusters of each Π_i.

Finally, we introduce an example using next generation sequencing datasets. Shen, Shen, Zhu, and Marron (2012, 2016) gave a scatter plot of first two PC scores for the next generation sequencing cancer data by Wilhelm and Landry (2009) in which the dataset consists of three curves with $d = 1{,}709$ and $n = 180$. See Figure 9 which was given in figure 1 of Shen et al. (2012). The three clusters seem to compose of a triangle such as Figure 7.



(a) Π_1, Π_2 and Π_3 (b) Π_1, Π_2 and Π_4 (c) Π_1, Π_3 and Π_4

FIGURE 7 Scatter plots of the first two principal component scores, supposing $k = 3$ in the dataset of Bhattacharjee et al. (2001). We denoted them by small circles when $\boldsymbol{x}_j \in \Pi_1$, by small triangles when $\boldsymbol{x}_j \in \Pi_2$, by small squares when $\boldsymbol{x}_j \in \Pi_3$, and by small inverted triangles when $\boldsymbol{x}_j \in \Pi_4$. The theoretical convergent points are denoted by the vertices of the triangle [Colour figure can be viewed at wileyonlinelibrary.com]



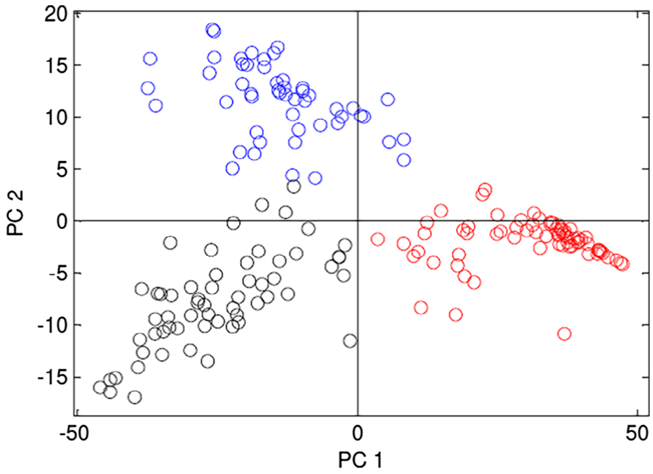
(i) $(\hat{z}_{1j}, \hat{z}_{2j})$

(ii) $(\hat{z}_{1j}, \hat{z}_{3j})$

(iii) $(\hat{z}_{1j}, \hat{z}_{2j}, \hat{z}_{3j})$

FIGURE 8 Scatter plots of the first three principal component scores, supposing $k = 4$ in the dataset of Bhattacharjee et al. (2001). The dashed triangles and triangular pyramid were given by Equation (7) with $k = 4$ [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 9 Shen et al. (2012) gave a scatter plot of first two principal component scores for the next generation sequencing cancer data. [Colour figure can be viewed at wileyonlinelibrary.com]



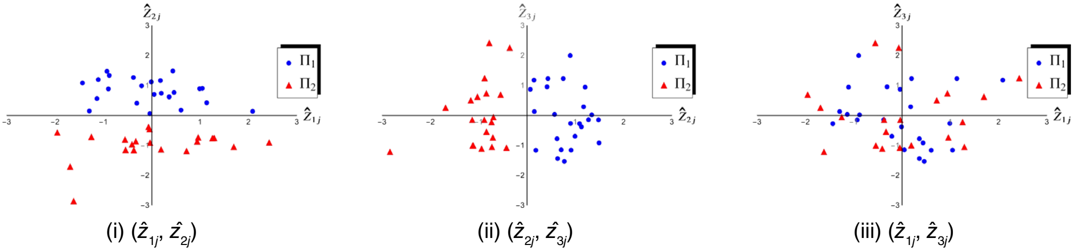
4.3 | Clustering: Special case

We analyzed microarray data by Armstrong et al. (2002) in which the dataset consists of three leukemia subtypes having 12,582 $(=d)$ genes. We used two classes such as Π_1: acute lymphoblastic leukemia (24 samples) and Π_2: mixed-lineage leukemia (20 samples), so that $n_1 = 24, n_2 = 20$, and $n = 44$. In Figure 10, we displayed scatter plots of the first three PC scores.

We observed that the dataset is perfectly separated by the sign of the second PC scores. This figure looks completely different from Figure 6. This is probably because the largest eigenvalue, $\lambda_{\max}(\boldsymbol{\Sigma}_1)$ or $\lambda_{\max}(\boldsymbol{\Sigma}_2)$, is too large. When $k = 2$, we give the following result to explain the reason of the phenomenon in Figure 10. Under the assumptions of Proposition 2, one can cluster $\boldsymbol{x}_j$s into two groups by some ith PC score even when Condition 1 is not met:

Proposition 2. *Assume that* $\max_{i=1,2}(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_1)/\Delta_1^2 \to 0$ *as* $d \to \infty$. *Then, there exists some positive integer* $i_\star$ *such that*

$$\frac{\lambda_{i_\star}}{\varepsilon_1\varepsilon_2\Delta_1} \to 1 \quad \text{as } d \to \infty.$$



(i) $(\hat{z}_{1j}, \hat{z}_{2j})$ (ii) $(\hat{z}_{2j}, \hat{z}_{3j})$ (iii) $(\hat{z}_{1j}, \hat{z}_{3j})$

FIGURE 10 Scatter plots of the first three principal component scores, supposing $k = 2$ in the dataset of Armstrong et al. (2002) [Colour figure can be viewed at wileyonlinelibrary.com]

Furthermore, assume that $\lambda_{i_\star}$ is distinct in the sense that

$$\liminf_{d\to\infty}\left|\frac{\lambda_{i'}}{\lambda_{i_\star}} - 1\right| > 0 \quad \text{for } i' = 1, \dots, d \ (i' \neq i_\star).$$

Then, if $\boldsymbol{h}_{i_\star}^{\mathrm{T}}\boldsymbol{\mu}_1 \geq 0$, it holds that Angle$(\boldsymbol{h}_{i_\star}, \boldsymbol{\mu}_1) \to 0$ as $d \to \infty$ and for $j = 1, \dots, n$

$$\operatorname*{plim}_{d\to\infty}\frac{s_{i_\star j}}{\lambda_{i_\star}^{1/2}} = \begin{cases} (\varepsilon_2/\varepsilon_1)^{1/2} & \textit{when } \boldsymbol{x}_j \in \Pi_1, \\ -(\varepsilon_1/\varepsilon_2)^{1/2} & \textit{when } \boldsymbol{x}_j \in \Pi_2. \end{cases}$$

We estimated the largest eigenvalue using the noise-reduction methodology given by Yata and Aoshima (2012). By noting Remark 1, we considered Δ_1 as $\Delta_1 = ||\boldsymbol{\mu}_1'||^2 = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$. We estimated Δ_1 using an unbiased estimator given by Aoshima and Yata (2014). Then, we obtained the estimates of $(\lambda_{\max}(\boldsymbol{\Sigma}_1)/\Delta_1, \lambda_{\max}(\boldsymbol{\Sigma}_2)/\Delta_1)$ as $(0.465, 0.787)$, so that Condition 1 is not met obviously. In addition, by estimating ε_is by η_is, we had $\varepsilon_2\lambda_{\max}(\boldsymbol{\Sigma}_2) > \varepsilon_1\varepsilon_2\Delta_1$. Thus, the first eigenspace of $\boldsymbol{\Sigma}$ is probably the first eigenspace of $\boldsymbol{\Sigma}_2$ as $\boldsymbol{\Sigma} = \varepsilon_1\varepsilon_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} + \varepsilon_1\boldsymbol{\Sigma}_1 + \varepsilon_2\boldsymbol{\Sigma}_2$. Thus, $i_\star$ in Proposition 2 must be 2. This is the reason why the dataset can be separated by the sign of the second PC scores in Figure 10.

5 | CONCLUDING REMARKS

In this article, we considered the mixture model by Equation (1) in high-dimensional settings. We studied asymptotic properties of both the true PC scores and the sample PC scores for the high-dimensional mixture model. We gave conditions under which PCA is very effective for clustering HDLSS data. We showed that HDLSS data can be classified by the sign of the first several PC scores theoretically. However, we have to say, in actual HDLSS data analyses, one may encounter cases such as in Figures 6c and 10, where the dataset is not always classified by the sign of the first several PC scores. Several reasons should be considered: (i) actual HDLSS datasets often include several outliers, (ii) the regularity conditions are not met, and (iii) the mixing proportions ε_is are quite unbalanced. Thus, we recommend the following three steps: (i) apply PCA to HDLSS data; (ii) using PC scores, map the dataset onto a feature space such as the first three eigenspaces, and (iii) apply general clustering methods such as the k-means method to the feature space. However, the number of clusters k is unknown in general. We emphasize that the first $k - 1$ eigenvalues are quite spiked for the model (1). Recently, Jung, Lee, and Ahn (2018) proposed a test of the number of spiked components for high-dimensional data. Thus, one may apply the test to the choice of k for clustering.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous referees for their constructive comments.

ORCID

Makoto Aoshima https://orcid.org/0000-0002-3791-7977

REFERENCES

Ahn, J., Lee, M. H., & Yoon, Y. J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, *22*, 443–464.

Ahn, J., Marron, J. S., Muller, K. M., & Chi, Y. Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, *94*, 760–766.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., & Marron, J. S. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, *60*, 4–19.

Aoshima, M., & Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis*, *30*, 356–399.

Aoshima, M., & Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, *66*, 983–1010.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., … Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, *30*, 41–47.

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

Borysov, P., Hannig, J., & Marron, J. S. (2014). Asymptotics of hierarchical clustering for growing dimension. *Journal of Multivariate Analysis*, *124*, 465–479.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., … Foa, R. (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.

Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B*, *67*, 427–444.

Hellton, K. H., & Thoresen, M. (2017). When and why are principal component scores a good tool for visualizing high-dimensional data? *Scandinavian Journal of Statistics*, *44*, 581–597.

Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, *7*, 359.

Jolliffe, I. T. (2002). *Principal component analysis*. New York, NY: Springer.

Jung, S., Lee, M. H., & Ahn, J. (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. *The Annals of Statistics*, *37*, 4104–4130.

Li, W., & Yao, J. (2018). On structure testing for component covariance matrices of a high dimensional mixture. *Journal of the Royal Statistical Society, Series B*, *80*, 293–318.

Liu, Y., Hayes, D. N., Nobel, A., & Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, *103*, 1281–1293.

Lv, J. (2013). Impacts of high dimensionality in finite samples. *The Annals of Statistics*, *41*, 2236–2262.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., … Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, *415*, 436–442.

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

Shen, D., Shen, H., Zhu, H., & Marron, J. S. (2012). *High dimensional principal component scores and data visualization*. arXiv:1211.2679.

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

Wilhelm, B. T., & Landry, J.-R. (2009). RNA-seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, *48*, 249–257.

Yata, K., & Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, *101*, 2060–2077.

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

How to cite this article: Yata K, Aoshima M. Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scand J Statist*. 2020;47:899–921. https://doi.org/10.1111/sjos.12432

APPENDIX A. LEMMAS AND THEIR PROOFS

Throughout, let $\boldsymbol{\mu}_{i,j} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and $\Delta_{i,j} = ||\boldsymbol{\mu}_{i,j}||^2$ for $i,j = 1, \dots, k (i < j)$. Let $\boldsymbol{u}_i = (u_{i1}, \dots, u_{in})^{\mathrm{T}}$, where

$$u_{ij} = \begin{cases} 0 & \text{when } i \geq 2 \text{ and } \boldsymbol{x}_j \in \overset{i-1}{\underset{m=1}{\cup}} \Pi_m, \\ [(1-\eta_{(i)})/\{n\eta_i(1-\eta_{(i-1)})\}]^{1/2} & \text{when } \boldsymbol{x}_j \in \Pi_i, \\ -[\eta_i/\{n(1-\eta_{(i)})(1-\eta_{(i-1)})\}]^{1/2} & \text{when } \boldsymbol{x}_j \in \overset{k}{\underset{m=i+1}{\cup}} \Pi_m, \end{cases}$$

for $i = 1, \dots, k-1; j = 1, \dots, n$. Let $\boldsymbol{v}_i = \sum_{m=1}^{k} \eta_m(\boldsymbol{\mu}_i - \boldsymbol{\mu}_m)$ for $i = 1, \dots, k$. Let $\boldsymbol{V} = [\boldsymbol{v}_{(1)}, \dots, \boldsymbol{v}_{(n)}]$, where $\boldsymbol{v}_{(j)} = \boldsymbol{v}_i$ according to $\boldsymbol{x}_j \in \Pi_i$ for $j = 1, \dots, n$. Note that $\boldsymbol{V}\boldsymbol{1}_n = \sum_{j=1}^{n} \boldsymbol{v}_{(j)} = \boldsymbol{0}$. We define the eigendecomposition of $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}/n$ by $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}/n = \sum_{i=1}^{k-1} \tilde{\lambda}_i \tilde{\boldsymbol{u}}_i \tilde{\boldsymbol{u}}_i^{\mathrm{T}}$ from the fact that $\mathrm{rank}(\boldsymbol{V}) \leq k-1$, where $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_{k-1} \geq 0$ are eigenvalues of $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}/n$ and $\tilde{\boldsymbol{u}}_i = (\tilde{u}_{i1}, \dots, \tilde{u}_{in})^{\mathrm{T}}$ is a unit eigenvector corresponding to $\tilde{\lambda}_i$ for each i. We assume $\tilde{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{u}_i \geq 0$ for $i = 1, \dots, k-1$, without loss of generality.

Lemma 1. *When $k = 2$, it holds that under Conditions 2–4*

$$\operatorname*{plim}_{d\to\infty} \frac{(n-1)\boldsymbol{S}_{\mathrm{D}} - tr(\boldsymbol{\Sigma}_1)\boldsymbol{P}_n}{\Delta_1} = \boldsymbol{r}\boldsymbol{r}^{\mathrm{T}}.$$

Proof. As $\boldsymbol{\mu}_2 = \boldsymbol{0}$, we can write that $\boldsymbol{x}_j - \eta_1\boldsymbol{\mu}_1 = (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + (-1)^{i+1}(1-\eta_i)\boldsymbol{\mu}_1$ for $i = 1, 2$; $j = 1, \dots, n$. From the fact that $\lambda_{\max}(\boldsymbol{\Sigma}_i) \leq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$, we have that $\mathrm{Var}\{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\mu}_1 | \boldsymbol{x}_j \in \Pi_i\} = \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_1 \leq \Delta_1\lambda_{\max}(\boldsymbol{\Sigma}_i) = o(\Delta_1^2)$ as $d \to \infty$ for $j = 1, \dots, n; i = 1, 2$ under Condition 2. Also, we have that $\mathrm{Var}\{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_{i'}) | \boldsymbol{x}_j \in \Pi_i, \boldsymbol{x}_{j'} \in \Pi_{i'}\} = \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'}) \leq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}\mathrm{tr}(\boldsymbol{\Sigma}_{i'}^2)^{1/2} = o(\Delta_1^2)$ for all $j \neq j'$ and $i, i' = 1, 2$ under Condition 2. Then, using Chebyshev's inequality, for any $\tau > 0$, under Condition 2, it holds that for all $j \neq j'$ and $i, i' = 1, 2$

$$\begin{aligned} &P\{|(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_{i'})/\Delta_1| > \tau | \boldsymbol{x}_j \in \Pi_i, \boldsymbol{x}_{j'} \in \Pi_{i'}\} = o(1) \text{ and} \\ &P\{|(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\mu}_1/\Delta_1| > \tau | \boldsymbol{x}_j \in \Pi_i\} = o(1), \end{aligned} \tag{A1}$$

so that $(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_{i'})/\Delta_1 = o_P(1)$ and $(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\mu}_1/\Delta_1 = o_P(1)$ when $\boldsymbol{x}_j \in \Pi_i$ and $\boldsymbol{x}_{j'} \in \Pi_{i'}$ $(j \neq j^{'})$. We note that $E(||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2 | \boldsymbol{x}_j \in \Pi_i) = \mathrm{tr}(\boldsymbol{\Sigma}_i)$. Similar to (A1), under Condition 3, it holds that $\{||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)\}/\Delta_1 = o_P(1)$ when $\boldsymbol{x}_j \in \Pi_i$ for $i = 1, 2; j = 1, \dots, n$. By noting that $\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\Delta_1 = o(1)$ under Condition 4, we have that

$$\operatorname*{plim}_{d \to \infty} \frac{(\boldsymbol{X} - \eta_1 \boldsymbol{\mu}_1 \boldsymbol{1}_n^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{X} - \eta_1 \boldsymbol{\mu}_1 \boldsymbol{1}_n^{\mathrm{T}}) - \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_n}{\Delta_1} = \boldsymbol{r}\boldsymbol{r}^{\mathrm{T}},$$

under Conditions 2–4. By noting that $\boldsymbol{P}_n(\boldsymbol{X} - \eta_1 \boldsymbol{\mu}_1 \boldsymbol{1}_n^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{X} - \eta_1 \boldsymbol{\mu}_1 \boldsymbol{1}_n^{\mathrm{T}})\boldsymbol{P}_n/(n-1) = \boldsymbol{S}_{\mathrm{D}}$ and $\boldsymbol{r}^{\mathrm{T}}\boldsymbol{P}_n = \boldsymbol{r}^{\mathrm{T}}$ from $\boldsymbol{r}^{\mathrm{T}}\boldsymbol{1}_n = 0$, we conclude the result. ■

Lemma 2. *Let* $\acute{\boldsymbol{\mu}}_{i,i+1} = \boldsymbol{\mu}_{i,i+1}/\Delta_{i,i+1}^{1/2}$ *for* $i = 1, \dots, k-1$ *and let* $\Delta_{(i,j)} = \Delta_{j,j+1}/\Delta_{i,i+1}$ *for* $i,j = 1, \dots, k-1 (i < j)$. *Under Conditions 1 and 5, it holds that as* $d \to \infty$

$$\frac{\lambda_i}{\Delta_{i,i+1}} = \frac{\varepsilon_i(1 - \varepsilon_{(i)})}{1 - \varepsilon_{(i-1)}} + o(1) \quad \textit{and} \quad \boldsymbol{h}_i^{\mathrm{T}} \acute{\boldsymbol{\mu}}_{i,i+1} = 1 + o(1) \quad \textit{for } i = 1, \dots, k-1;$$

$$\boldsymbol{h}_i^{\mathrm{T}} \acute{\boldsymbol{\mu}}_{i-1,i} = -\frac{1 - \varepsilon_{(i)}}{1 - \varepsilon_{(i-1)}} \Delta_{(i-1,i)}^{1/2} \{1 + o(1)\} \quad \textit{for } i = 2, \dots, k-1 \textit{ when } k \geq 3; \quad \textit{and}$$

$$\boldsymbol{h}_j^{\mathrm{T}} \acute{\boldsymbol{\mu}}_{i,i+1} = o(\Delta_{(i,j)}^{1/2}) \quad \text{for } i,j = 1, \dots, k-1 (i+1 < j) \textit{ when } k \geq 3.$$

Proof. From the fact that $|\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\mu}_j| \leq (\Delta_i \Delta_j)^{1/2}$, under Condition 5 it holds that as $d \to \infty$

$$\frac{\Delta_{i,i+1}}{\Delta_i} = \frac{\Delta_i + \Delta_{i+1} + O\{(\Delta_i \Delta_{i+1})^{1/2}\}}{\Delta_i} \to 1 \quad \text{for } i = 1, \dots, k-2.$$

Then, under Condition 5, it holds that

$$\frac{\boldsymbol{\mu}_{i,i+1}^{\mathrm{T}} \boldsymbol{\mu}_{j,j+1}}{(\Delta_{i,i+1}\Delta_{j,j+1})^{1/2}} = \frac{\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\mu}_j + O\{(\Delta_i \Delta_{j+1})^{1/2} + (\Delta_{i+1}\Delta_j)^{1/2}\}}{(\Delta_i \Delta_j)^{1/2}\{1 + o(1)\}} = o(1),$$

for $i,j = 1, \dots, k-1 (i < j)$. Hence, under Condition 5, we claim that

$$\acute{\boldsymbol{\mu}}_{i,i+1}^{\mathrm{T}} \acute{\boldsymbol{\mu}}_{j,j+1} \to 0 \quad \text{and} \quad \frac{\Delta_{j,j+1}}{\Delta_{i,i+1}} \to 0 \text{ as } d \to \infty \text{ for } i,j = 1, \dots, k-1 \ (i < j). \tag{A2}$$

Let $\boldsymbol{e}_d$ $(\in \mathbb{R}^d)$ be an arbitrary unit vector. From $\boldsymbol{e}_d^{\mathrm{T}} \left(\sum_{i=1}^k \varepsilon_i \boldsymbol{\Sigma}_i \right) \boldsymbol{e}_d \leq \sum_{i=1}^k \lambda_{\max}(\boldsymbol{\Sigma}_i)$, it holds that

$$\frac{\boldsymbol{e}_d^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{e}_d}{\Delta_{k-1}} = \frac{\boldsymbol{e}_d^{\mathrm{T}} \left(\sum_{i=1}^{k-1} \sum_{j=i+1}^k \varepsilon_i \varepsilon_j \boldsymbol{\mu}_{i,j} \boldsymbol{\mu}_{i,j}^{\mathrm{T}} \right) \boldsymbol{e}_d}{\Delta_{k-1}} + o(1), \tag{A3}$$

under Condition 1. Note that $\boldsymbol{\mu}_{i,j} = \sum_{m=i}^{j-1} \boldsymbol{\mu}_{m,m+1}$ for $i,j = 1, \dots, k (i < j)$. Thus, it holds that

$$\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \varepsilon_i \varepsilon_j \boldsymbol{\mu}_{i,j} \boldsymbol{\mu}_{i,j}^{\mathrm{T}} = \sum_{i=1}^{k-1} \varepsilon_{(i)}(1 - \varepsilon_{(i)}) \boldsymbol{\mu}_{i,i+1} \boldsymbol{\mu}_{i,i+1}^{\mathrm{T}} + \sum_{i=1}^{k-2} \sum_{j=i+1}^{k-1} \varepsilon_{(i)}(1 - \varepsilon_{(j)})(\boldsymbol{\mu}_{i,i+1} \boldsymbol{\mu}_{j,j+1}^{\mathrm{T}} + \boldsymbol{\mu}_{j,j+1} \boldsymbol{\mu}_{i,i+1}^{\mathrm{T}}). \tag{A4}$$

From the facts that $\lambda_1 = \boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{h}_1 = \max_{\boldsymbol{e}_d}(\boldsymbol{e}_d^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{e}_d)$ and $\Delta_{k-1} = \Delta_{k-1,k}$, by combining Equation (A3) with Equations (A2) and (A4), we have that

$$\frac{\lambda_1}{\Delta_{1,2}} = \max_{\boldsymbol{e}_d}\{\varepsilon_{(1)}(1-\varepsilon_{(1)})(\boldsymbol{e}_d^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{1,2})^2 + o(1)\} = \varepsilon_{(1)}(1-\varepsilon_{(1)}) + o(1),$$

under Conditions 1 and 5. Hence, by noting that $\Delta_{1,2}/\Delta_1 = 1 + o(1)$ and $\boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\mu}_1 \geq 0$, it holds that $\boldsymbol{h}_1^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{1,2} = \boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\mu}_1/\Delta_1^{1/2} + o(1) = 1 + o(1)$.

Next, we consider λ_2 and $\boldsymbol{h}_2$. From (A2), we note that $\acute{\boldsymbol{\mu}}_{i,i+1}^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{j,j+1} = o(1)$ and $\Delta_{(i,j)} = o(1)$ for $i,j = 1,\ldots,k-1(i<j)$ under Condition 5. Then, under Conditions 1 and 5, it holds that for $j \geq 2$

$$0 = \frac{\boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{h}_j}{\Delta_{1,2}} = \varepsilon_{(1)}(1-\varepsilon_{(1)})\{1+o(1)\}\acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}}\boldsymbol{h}_j + \varepsilon_{(1)}(1-\varepsilon_{(2)})\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(1,2)}^{1/2} + o(\Delta_{(1,2)}^{1/2}),$$

from Equations (A3) and (A4) and $\boldsymbol{h}_1^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{2,3} = o(1)$, so that for $j \geq 2$

$$\boldsymbol{h}_j^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{1,2} = -\{(1-\varepsilon_{(2)})/(1-\varepsilon_{(1)})\}\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(1,2)}^{1/2} + o(\Delta_{(1,2)}^{1/2}). \tag{A5}$$

By combining Equation (A3) with Equations (A4) and (A5), we have that

$$\begin{aligned}\frac{\lambda_2}{\Delta_{2,3}} = \frac{\boldsymbol{h}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{h}_2}{\Delta_{2,3}} &= \frac{\boldsymbol{h}_2^{\mathrm{T}}\left\{\sum_{i=1}^{2}\varepsilon_{(i)}(1-\varepsilon_{(i)})\boldsymbol{\mu}_{i,i+1}\boldsymbol{\mu}_{i,i+1}^{\mathrm{T}} + 2\varepsilon_{(1)}(1-\varepsilon_{(2)})\boldsymbol{\mu}_{1,2}\boldsymbol{\mu}_{2,3}^{\mathrm{T}}\right\}\boldsymbol{h}_2}{\Delta_{2,3}} + o(1)\\ &= \varepsilon_{(2)}(1-\varepsilon_{(2)})(\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_2)^2 + \varepsilon_{(1)}(1-\varepsilon_{(1)})\frac{(\acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}}\boldsymbol{h}_2)^2}{\Delta_{(1,2)}} + 2\varepsilon_{(1)}(1-\varepsilon_{(2)})\frac{(\acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}}\boldsymbol{h}_2)(\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_2)}{\Delta_{(1,2)}^{1/2}} + o(1)\\ &= \varepsilon_{(2)}(1-\varepsilon_{(2)}) - \frac{\varepsilon_{(1)}(1-\varepsilon_{(2)})^2}{1-\varepsilon_{(1)}} + o(1) = \frac{\varepsilon_2(1-\varepsilon_{(2)})}{1-\varepsilon_{(1)}} + o(1),\end{aligned} \tag{A6}$$

under Conditions 1 and 5. Hence, from the assumption that $\boldsymbol{h}_2^{\mathrm{T}}\boldsymbol{\mu}_2 \geq 0$, it holds that $\boldsymbol{h}_2^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{2,3} = \boldsymbol{h}_2^{\mathrm{T}}\boldsymbol{\mu}_2/\Delta_2^{1/2} + o(1) = 1 + o(1)$.

Next, we consider λ_3 and $\boldsymbol{h}_3$. Note that $\boldsymbol{h}_j^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{2,3} = o(1)$ for $j \geq 3$ from $\boldsymbol{h}_2^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{2,3} = 1 + o(1)$. Then, under Conditions 1 and 5, we have that for $j \geq 3$

$$\begin{aligned}0 = \frac{\boldsymbol{h}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{h}_j}{\Delta_{1,2}} &= \varepsilon_{(1)}(1-\varepsilon_{(1)})\{1+o(1)\}\acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}}\boldsymbol{h}_j + \varepsilon_{(1)}(1-\varepsilon_{(2)})\{1+o(1)\}\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(1,2)}^{1/2}\\ &\quad + \varepsilon_{(1)}(1-\varepsilon_{(3)})\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(1,3)}^{1/2} + o(\Delta_{(1,3)}^{1/2}) \quad \text{and}\end{aligned} \tag{A7}$$

$$\begin{aligned}0 = \frac{\boldsymbol{h}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{h}_j}{\Delta_{2,3}} &= \varepsilon_{(1)}(1-\varepsilon_{(1)})\frac{\boldsymbol{h}_2^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{1,2}\acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}}\boldsymbol{h}_j}{\Delta_{(1,2)}} + \varepsilon_{(1)}(1-\varepsilon_{(2)})\frac{\boldsymbol{h}_2^{\mathrm{T}}(\acute{\boldsymbol{\mu}}_{1,2}\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}} + \acute{\boldsymbol{\mu}}_{2,3}\acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}})\boldsymbol{h}_j}{\Delta_{(1,2)}^{1/2}}\\ &\quad + \varepsilon_{(1)}(1-\varepsilon_{(3)})\frac{\boldsymbol{h}_2^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{1,2}\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_j}{\Delta_{(1,2)}^{1/2}}\Delta_{(2,3)}^{1/2} + \varepsilon_{(2)}(1-\varepsilon_{(2)})\{1+o(1)\}\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_j\\ &\quad + \varepsilon_{(2)}(1-\varepsilon_{(3)})\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(2,3)}^{1/2} + o(\Delta_{(2,3)}^{1/2})\\ &= \frac{\varepsilon_2(1-\varepsilon_{(2)})}{1-\varepsilon_{(1)}}\{1+o(1)\}\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_j + \frac{\varepsilon_2(1-\varepsilon_{(3)})}{1-\varepsilon_{(1)}}\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(2,3)}^{1/2} + o(\Delta_{(2,3)}^{1/2}) + \acute{\boldsymbol{\mu}}_{1,2}^{\mathrm{T}}\boldsymbol{h}_j \times o(\Delta_{(1,2)}^{-1/2}),\end{aligned} \tag{A8}$$

from Equations (A2) to (A5), $\boldsymbol{h}_1^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{2,3}=o(1)$, $\boldsymbol{h}_1^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{3,4}=o(1)$, and $\boldsymbol{h}_2^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{3,4}=o(1)$. Then, by combining Equations (A7) and (A8), under Conditions 1 and 5, it holds that for $j\ge 3$

$$\boldsymbol{h}_j^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{1,2}=o(\Delta_{(1,3)}^{1/2}) \quad \text{and} \quad \boldsymbol{h}_j^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{2,3}=-\frac{1-\varepsilon_{(3)}}{1-\varepsilon_{(2)}}\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_j\Delta_{(2,3)}^{1/2}+o(\Delta_{(2,3)}^{1/2}). \tag{A9}$$

Similar to Equation (A6), by combining Equation (A3) with Equations (A4) and (A9), under Conditions 1 and 5, we have that

$$\begin{aligned}\frac{\lambda_3}{\Delta_{3,4}}&=\varepsilon_{(3)}(1-\varepsilon_{(3)})(\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_3)^2+\varepsilon_{(2)}(1-\varepsilon_{(2)})\frac{(\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_3)^2}{\Delta_{(2,3)}}+2\varepsilon_{(2)}(1-\varepsilon_{(3)})\frac{(\acute{\boldsymbol{\mu}}_{2,3}^{\mathrm{T}}\boldsymbol{h}_3)(\acute{\boldsymbol{\mu}}_{3,4}^{\mathrm{T}}\boldsymbol{h}_3)}{\Delta_{(2,3)}^{1/2}}+o(1)\\&=\varepsilon_{(3)}(1-\varepsilon_{(3)})-\frac{\varepsilon_{(2)}(1-\varepsilon_{(3)})^2}{1-\varepsilon_{(2)}}+o(1)=\frac{\varepsilon_3(1-\varepsilon_{(3)})}{(1-\varepsilon_{(2)})}+o(1),\end{aligned}$$

so that $\boldsymbol{h}_3^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{3,4}=1+o(1)$ from the assumption that $\boldsymbol{h}_3^{\mathrm{T}}\boldsymbol{\mu}_3\ge 0$.

In a way similar to λ_3 and $\boldsymbol{h}_3$, as for λ_i and $\boldsymbol{h}_i$ $(4\le i\le k-1)$, we have that $\lambda_i/\Delta_{i,i+1}=\varepsilon_i(1-\varepsilon_{(i)})/(1-\varepsilon_{(i-1)})+o(1)$, $\boldsymbol{h}_i^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{i,i+1}=1+o(1)$ and $\boldsymbol{h}_i^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{i-1,i}=-\{(1-\varepsilon_{(i)})/(1-\varepsilon_{(i-1)})\}\Delta_{(i-1,i)}^{1/2}\{1+o(1)\}$ together with $\boldsymbol{h}_j^{\mathrm{T}}\acute{\boldsymbol{\mu}}_{i,i+1}=o(\Delta_{(i,j)}^{1/2})$ for $i,j=1,\dots,k-1$ $(i+1<j)$ under Conditions 1 and 5. It concludes the results. ■

Lemma 3. *Under Conditions 1 and 5, it holds that for* $i=1,\dots,k-1$

$$\lim_{d\to\infty}\boldsymbol{h}_i^{\mathrm{T}}\sum_{m=1}^{k}\frac{\varepsilon_m(\boldsymbol{\mu}_{i'}-\boldsymbol{\mu}_m)}{\lambda_i^{1/2}}=\begin{cases}0 & \textit{when } i\ge 2 \textit{ and } i'<i,\\ [(1-\varepsilon_{(i)})/\{\varepsilon_i(1-\varepsilon_{(i-1)})\}]^{1/2} & \textit{when } i'=i,\\ -[\varepsilon_i/\{(1-\varepsilon_{(i)})(1-\varepsilon_{(i-1)})\}]^{1/2} & \textit{when } i'>i.\end{cases}$$

Proof. We write that

$$\begin{aligned}&\sum_{m=1}^{k}\varepsilon_m(\boldsymbol{\mu}_1-\boldsymbol{\mu}_m)=\sum_{m=1}^{k-1}(1-\varepsilon_{(m)})\boldsymbol{\mu}_{m,m+1},\\&\sum_{m=1}^{k}\varepsilon_m(\boldsymbol{\mu}_k-\boldsymbol{\mu}_m)=-\sum_{m=1}^{k-1}\varepsilon_{(m)}\boldsymbol{\mu}_{m,m+1}\quad\text{and}\\&\sum_{m=1}^{k}\varepsilon_m(\boldsymbol{\mu}_i-\boldsymbol{\mu}_m)=\sum_{m=i}^{k-1}(1-\varepsilon_{(m)})\boldsymbol{\mu}_{m,m+1}-\sum_{m=1}^{i-1}\varepsilon_{(m)}\boldsymbol{\mu}_{m,m+1},\end{aligned}\tag{A10}$$

for $i=2,\dots,k-1$. Using Lemma 2, under Conditions 1 and 5, we have that as $d\to\infty$

$$\begin{aligned}&\boldsymbol{h}_1^{\mathrm{T}}\sum_{m=1}^{k}\frac{\varepsilon_m(\boldsymbol{\mu}_1-\boldsymbol{\mu}_m)}{\Delta_{1,2}^{1/2}}=\boldsymbol{h}_1^{\mathrm{T}}\frac{(1-\varepsilon_{(1)})\boldsymbol{\mu}_{1,2}}{\Delta_{1,2}^{1/2}}+o(1)=1-\varepsilon_{(1)}+o(1)\quad\text{and}\\&\boldsymbol{h}_1^{\mathrm{T}}\sum_{m=1}^{k}\frac{\varepsilon_m(\boldsymbol{\mu}_{i'}-\boldsymbol{\mu}_m)}{\Delta_{i,i+1}^{1/2}}=-\boldsymbol{h}_1^{\mathrm{T}}\frac{\varepsilon_{(1)}\boldsymbol{\mu}_{1,2}}{\Delta_{1,2}^{1/2}}+o(1)=-\varepsilon_{(1)}+o(1)\quad\text{for } i'=2,\dots,k,\end{aligned}$$

from Equation (A10). Also, using Lemma 2, under Conditions 1 and 5, we have that for $i=2,\dots,k-1$; $i'=i+1,\dots,k$; $i''=1,\dots,i-1$

$$\boldsymbol{h}_i^{\mathrm{T}}\sum_{m=1}^{k}\frac{\varepsilon_m(\boldsymbol{\mu}_i-\boldsymbol{\mu}_m)}{\Delta_{i,i+1}^{1/2}}=\boldsymbol{h}_i^{\mathrm{T}}\frac{(1-\varepsilon_{(i)})\boldsymbol{\mu}_{i,i+1}-\varepsilon_{(i-1)}\boldsymbol{\mu}_{i-1,i}}{\Delta_{i,i+1}^{1/2}}+o(1)$$

$$
\begin{aligned}
&= (1-\varepsilon_{(i)}) + \frac{\varepsilon_{(i-1)}(1-\varepsilon_{(i)})}{1-\varepsilon_{(i-1)}} + o(1)\\
&= \frac{1-\varepsilon_{(i)}}{1-\varepsilon_{(i-1)}} + o(1),\\
\boldsymbol{h}_i^{\mathrm{T}} \sum_{m=1}^{k} \frac{\varepsilon_m(\boldsymbol{\mu}_{i'}-\boldsymbol{\mu}_m)}{\Delta_{i,i+1}^{1/2}} &= -\varepsilon_{(i)} + \frac{\varepsilon_{(i-1)}(1-\varepsilon_{(i)})}{1-\varepsilon_{(i-1)}} + o(1)\\
&= -\frac{\varepsilon_i}{1-\varepsilon_{(i-1)}} + o(1), \quad \text{and}\\
\boldsymbol{h}_i^{\mathrm{T}} \sum_{m=1}^{k} \frac{\varepsilon_m(\boldsymbol{\mu}_{i''}-\boldsymbol{\mu}_m)}{\Delta_{i,i+1}^{1/2}} &= o(1).
\end{aligned}
$$

Thus, from Lemma 2, we can conclude the results. ■

Lemma 4. *Assume Conditions 2–6. Then, under the condition:*

$$
0 < \operatorname*{plim}_{d\to\infty} \frac{\tilde{\lambda}_i}{\Delta_{i,i+1}} < \infty \quad \textit{for } i = 1, \dots, k-1, \tag{A11}
$$

it holds that

$$
\operatorname*{plim}_{d\to\infty} \hat{\boldsymbol{u}}_i^{\mathrm{T}} \tilde{\boldsymbol{u}}_i = 1 \quad \textit{for } \hat{\boldsymbol{u}}_i^{\mathrm{T}} \tilde{\boldsymbol{u}}_i \geq 0,\ i = 1, \dots, k-1.
$$

Proof. From the fact that $\lambda_{\max}(\boldsymbol{\Sigma}_i) \leq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$, we have that $\mathrm{Var}\{\boldsymbol{\mu}_{k-1}^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu}_i) | \boldsymbol{x}_j \in \Pi_i\} = \boldsymbol{\mu}_{k-1}^{\mathrm{T}} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_{k-1} \leq \lambda_{\max}(\boldsymbol{\Sigma}_i)\Delta_{k-1} = o(\Delta_{k-1}^2)$ as $d \to \infty$ for $i = 1, \dots, k; j = 1, \dots, n$ under Condition 2. Then, we have that for $i = 1, \dots, k-1;\ i' = 1, \dots, k; j = 1, \dots, n$

$$
\mathrm{Var}\{\boldsymbol{\mu}_{i,i+1}^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu}_{i'}) | \boldsymbol{x}_j \in \Pi_{i'}\} = \boldsymbol{\mu}_{i,i+1}^{\mathrm{T}} \boldsymbol{\Sigma}_{i'} \boldsymbol{\mu}_{i,i+1} = O(\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{i'} \boldsymbol{\mu}_i + \boldsymbol{\mu}_{i+1}^{\mathrm{T}} \boldsymbol{\Sigma}_{i'} \boldsymbol{\mu}_{i+1}) = o(\Delta_{k-1}^2),
$$

under Conditions 2 and 6. Then, similar to Equation (A1), under Conditions 2 and 6, it holds that $\boldsymbol{\mu}_{i,i+1}^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu}_{i'})/\Delta_{k-1} = o_P(1)$ when $\boldsymbol{x}_j \in \Pi_{i'}$ for $i = 1, \dots, k-1; i' = 1, \dots, k; j = 1, \dots, n$. In addition, under Conditions 2 and 3, we can claim that $(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_{i'})/\Delta_{k-1} = o_P(1)$ and $||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2/\Delta_{k-1} = \mathrm{tr}(\boldsymbol{\Sigma}_i)/\Delta_{k-1} + o_P(1)$ when $\boldsymbol{x}_j \in \Pi_i$ and $\boldsymbol{x}_{j'} \in \Pi_{i'}$ for all $j \neq j'$ and $i, i' = 1, \dots, k$. Here, we write that $\boldsymbol{x}_j - \boldsymbol{\mu}_\eta = (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \boldsymbol{v}_i$ for $i = 1, \dots, k;\ j = 1, \dots, n$, where $\boldsymbol{\mu}_\eta = \sum_{i=1}^{k} \eta_i \boldsymbol{\mu}_i$. Then, by noting Equation (A10) with $\varepsilon_i = \eta_i$ and $\varepsilon_{(i)} = \eta_{(i)}$, $i = 1, \dots, k$, under Conditions 2, 3, and 6, we have that

$$
\frac{||\boldsymbol{x}_j - \boldsymbol{\mu}_\eta||^2}{\Delta_{k-1}} = \frac{||\boldsymbol{v}_i||^2 + \mathrm{tr}(\boldsymbol{\Sigma}_i)}{\Delta_{k-1}} + o_P(1) \quad \text{and} \quad \frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_\eta)^{\mathrm{T}}(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_\eta)}{\Delta_{k-1}} = \frac{\boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{v}_{i'}}{\Delta_{k-1}} + o_P(1),
$$

when $\boldsymbol{x}_j \in \Pi_i$ and $\boldsymbol{x}_{j'} \in \Pi_{i'}$ for all $j \neq j'$ and $i, i' = 1, \dots, k$. Thus, under Conditions 2, 3, 4, and 6, it holds that

$$
\operatorname*{plim}_{d\to\infty} \frac{(\boldsymbol{X} - \boldsymbol{\mu}_\eta \boldsymbol{1}_n^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\mu}_\eta \boldsymbol{1}_n^{\mathrm{T}}) - \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_n - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}{\Delta_{k-1}} = \boldsymbol{O}. \tag{A12}
$$

Let $\boldsymbol{e}_{n*}$ $(\in \mathbb{R}^n)$ be an arbitrary random unit vector such that $\boldsymbol{e}_{n*}^{\mathrm{T}}\boldsymbol{1}_n = 0$. We note that $\boldsymbol{P}_n(\boldsymbol{X} - \boldsymbol{\mu}_\eta \boldsymbol{1}_n^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\mu}_\eta \boldsymbol{1}_n^{\mathrm{T}})\boldsymbol{P}_n/(n-1) = \boldsymbol{S}_{\mathrm{D}}$. Then, by noting $\boldsymbol{e}_{n*}^{\mathrm{T}}\boldsymbol{P}_n = \boldsymbol{e}_{n*}^{\mathrm{T}}$, under Equation (A11), Conditions 2, 3, 4, and 6, we have that

$$\begin{aligned}
\boldsymbol{e}_{n*}^{\mathrm{T}}\frac{(n-1)\boldsymbol{S}_{\mathrm{D}} - \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{P}_n}{\Delta_{k-1}}\boldsymbol{e}_{n*} &= \frac{\sum_{i=1}^{n-1}(n-1)\hat{\lambda}_i \boldsymbol{e}_{n*}^{\mathrm{T}}\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{e}_{n*} - \mathrm{tr}(\boldsymbol{\Sigma}_1)}{\Delta_{k-1}} \\
&= \boldsymbol{e}_{n*}^{\mathrm{T}}\frac{(\boldsymbol{X} - \boldsymbol{\mu}_\eta \boldsymbol{1}_n^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{\mu}_\eta \boldsymbol{1}_n^{\mathrm{T}}) - \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_n}{\Delta_{k-1}}\boldsymbol{e}_{n*} \\
&= \boldsymbol{e}_{n*}^{\mathrm{T}}\frac{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}{\Delta_{k-1}}\boldsymbol{e}_{n*} + o_P(1) \\
&= \frac{\sum_{i=1}^{k-1} n\tilde{\lambda}_i \boldsymbol{e}_{n*}^{\mathrm{T}}\tilde{\boldsymbol{u}}_i\tilde{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{e}_{n*}}{\Delta_{k-1}} + o_P(1), \qquad \text{(A13)}
\end{aligned}$$

from Equation (A12). We note that $\tilde{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{1}_n = 0$ for $i = 1, \dots, k-1$ in case of $\mathrm{rank}(\boldsymbol{V}) = k-1$. Also, from Equation (A2), we note that $\tilde{\lambda}_i$, $i = 1, \dots, k-1$, are distinct under Condition 5 and Equation (A11) for a sufficiently large d. Thus, from Equation (A13), if $\hat{\boldsymbol{u}}_i^{\mathrm{T}}\tilde{\boldsymbol{u}}_i \geq 0$ for $i = 1, \dots, k-1$, we have that $\hat{\boldsymbol{u}}_i^{\mathrm{T}}\tilde{\boldsymbol{u}}_i = 1 + o_P(1)$ for $i = 1, \dots, k-1$. It concludes the result. ■

Lemma 5. *Assume Condition 5. For $n_i > 0, i = 1, \dots, k$, it holds that*

$$\operatorname*{plim}_{d\to\infty}\frac{\tilde{\lambda}_i}{\Delta_{i,i+1}} = \frac{\eta_i(1-\eta_{(i)})}{1-\eta_{(i-1)}} \quad \textit{and} \quad \operatorname*{plim}_{d\to\infty}\tilde{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{u}_i = 1 \quad \textit{for } i = 1, \dots, k-1.$$

Proof. By noting Equation (A10) with $\varepsilon_i = \eta_i$ and $\varepsilon_{(i)} = \eta_{(i)}$, $i = 1, \dots, k$, we can write that

$$\frac{\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}}{n} = \sum_{i=1}^{k-1}\eta_{(i)}(1-\eta_{(i)})\boldsymbol{\mu}_{i,i+1}\boldsymbol{\mu}_{i,i+1}^{\mathrm{T}} + \sum_{i=1}^{k-2}\sum_{j=i+1}^{k-1}\eta_{(i)}(1-\eta_{(j)})(\boldsymbol{\mu}_{i,i+1}\boldsymbol{\mu}_{j,j+1}^{\mathrm{T}} + \boldsymbol{\mu}_{j,j+1}\boldsymbol{\mu}_{i,i+1}^{\mathrm{T}}). \qquad \text{(A14)}$$

We have the eigendecomposition of $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}/n$ by $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}/n = \sum_{i=1}^{k-1}\tilde{\lambda}_i\tilde{\boldsymbol{h}}_i\tilde{\boldsymbol{h}}_i^{\mathrm{T}}$, where $\tilde{\boldsymbol{h}}_i$ is a unit eigenvector corresponding to $\tilde{\lambda}_i$ for each i. We note that $\eta_i > 0, i = 1, \dots, k$ for $n_i > 0$, $i = 1, \dots, k$. Then, by noting Lemmas 2 and 3 and the fact that Equation (A14) is same as Equation (A4) with $\varepsilon_{(i)} = \eta_{(i)}$, $i = 1, \dots, k-1$, under Condition 5, we have that for $i = 1, \dots, k-1$

$$\operatorname*{plim}_{d\to\infty}\frac{\tilde{\lambda}_i}{\Delta_{i,i+1}} = \frac{\eta_i(1-\eta_{(i)})}{1-\eta_{(i-1)}} \quad \text{and} \quad \operatorname*{plim}_{d\to\infty}\frac{\tilde{\boldsymbol{h}}_i^{\mathrm{T}}\boldsymbol{v}_{(j)}}{\tilde{\lambda}_i^{1/2}} = u_{ij}n^{1/2},$$

if $\tilde{\boldsymbol{h}}_i^{\mathrm{T}}\boldsymbol{\mu}_i \geq 0$. We note that $\tilde{u}_{ij} = \tilde{\boldsymbol{h}}_i^{\mathrm{T}}\boldsymbol{v}_{(j)}/(n\tilde{\lambda}_i)^{1/2}$ from the fact that $\tilde{\boldsymbol{u}}_i = \boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{h}}_i/(n\tilde{\lambda}_i)^{1/2}$ for $i = 1, \dots, k-1$. Hence, we can conclude the result. ■

APPENDIX B. ADDITIONAL PROPOSITION

When Condition 5 is not met, Theorem 3 does not hold. However, in Figure 5c, we could find three separate clusters of Π_i, $i = 1, 2, 3$, even though Condition 5 is not met. To explain the reason of this phenomenon, we give the following result.

Proposition 3. *Assume Conditions* 2–4 *and* 6. *Then, under the condition:*

$$0 < \operatorname*{plim}_{d\to\infty} \frac{\tilde{\lambda}_{k-1}}{\Delta_{k-1}} < \infty,$$

it holds that for $i = 1, \dots, k-1$, *as* $d \to \infty$

$$\hat{\boldsymbol{u}}_i^{\mathrm{T}} \frac{(n-1)\boldsymbol{S}_{\mathrm{D}}}{\Delta_{k-1}} \hat{\boldsymbol{u}}_i = \frac{tr(\boldsymbol{\Sigma}_1)}{\Delta_{k-1}} + \hat{\boldsymbol{u}}_i^{\mathrm{T}} \frac{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}{\Delta_{k-1}} \hat{\boldsymbol{u}}_i + o_P(1).$$

Proof. By noting that $\hat{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{1}_n = 0$ for $i = 1, \dots, k-1$ when $\operatorname{rank}(\boldsymbol{S}_{\mathrm{D}}) \ge k-1$, from Equation (A13), we can conclude the result. ■

By noting that $\hat{\boldsymbol{u}}_i = (\hat{z}_{i1}, \dots, \hat{z}_{in})^{\mathrm{T}}/n^{1/2}$, from Proposition 3, for sufficiently large d, the estimated PC scores depend only on the structure of $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$ even when Condition 5 is not met. Then, as $\operatorname{rank}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}) = k-1$, there must be k separate clusters for $\Pi_i, i = 1, \dots, k$, in the first $k-1$ PC spaces as seen in Figure 5c.

APPENDIX C. PROOFS OF THEOREMS, COROLLARIES, AND PROPOSITIONS

Proofs of Theorem 1 and Corollary 1. We note that $\operatorname{tr}(\boldsymbol{\Sigma}_1)/\operatorname{tr}(\boldsymbol{\Sigma}) \to (1 - \varepsilon_1\varepsilon_2 c)$ as $d \to \infty$ under Condition 4 and $\Delta_1/\operatorname{tr}(\boldsymbol{\Sigma}) \to c(> 0)$ as $d \to \infty$. Then, using Lemma 1, we can conclude the result of Theorem 1.

Next, we consider the proof of Corollary 1. From the fact that $\boldsymbol{1}_n^{\mathrm{T}}\boldsymbol{S}_{\mathrm{D}}\boldsymbol{1}_n = 0$, it holds that $\hat{\boldsymbol{u}}_1^{\mathrm{T}}\boldsymbol{1}_n = 0$ when $\boldsymbol{S}_{\mathrm{D}} \ne \boldsymbol{O}$, so that $\boldsymbol{P}_n\hat{\boldsymbol{u}}_1 = \hat{\boldsymbol{u}}_1$. Also, note that $||\boldsymbol{r}||^2 = n\eta_1\eta_2$ and $\boldsymbol{r}^{\mathrm{T}}\boldsymbol{1}_n = 0$. Then, using Lemma 1, under Conditions 2–4, it holds that $\hat{\boldsymbol{u}}_1^{\mathrm{T}}\{(n-1)\boldsymbol{S}_{\mathrm{D}} - \operatorname{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{P}_n\}\hat{\boldsymbol{u}}_1/\Delta_1 = n\eta_1\eta_2 + o_P(1)$. Hence, from Equation (3) and the assumption that $\hat{\boldsymbol{u}}_1^{\mathrm{T}}\boldsymbol{z}_1 \ge 0$, we have that $\hat{\boldsymbol{u}}_1^{\mathrm{T}}\{(n\eta_1\eta_2)^{-1/2}\boldsymbol{r}\} = 1 + o_P(1)$ for $n_i > 0$, $i = 1, 2$. In view of the elements of $\boldsymbol{r}$, we can conclude the result of Corollary 1. ■

Proof of Proposition 1. We assume $\boldsymbol{x}_j \in \Pi_1$ for $j = 1, \dots, n_1$, $\boldsymbol{x}_j \in \Pi_2$ for $j = n_1 + 1, \dots, n$, and $\operatorname{tr}(\boldsymbol{\Sigma}_1) \ge \operatorname{tr}(\boldsymbol{\Sigma}_2)$ without loss of generality. Similar to the proof of Lemma 1, under the assumptions of Proposition 1, we have that

$$\operatorname*{plim}_{d\to\infty} \frac{(n-1)\boldsymbol{S}_{\mathrm{D}} - \operatorname{tr}(\boldsymbol{\Sigma}_2)\boldsymbol{P}_n}{\Delta_{\Sigma}} = \boldsymbol{P}_n\boldsymbol{D}_n\boldsymbol{P}_n,$$

where $\boldsymbol{D}_n = \operatorname{diag}(1, \dots, 1, 0, \dots, 0)$ whose first n_1 diagonal elements are 1. Note that the first $n_1 - 1$ eigenvalues of $\boldsymbol{P}_n\boldsymbol{D}_n\boldsymbol{P}_n$ are multiple. Also, note that the eigenspace for the multiple eigenvalue consists of the $n_1 - 1$ vectors,

$$(1, -1, 0, \dots, 0)^{\mathrm{T}}, (1, 0, -1, 0, \dots, 0)^{\mathrm{T}}, \dots, (1, 0, \dots, 0, -1, 0, \dots, 0)^{\mathrm{T}}.$$

Thus, by noting that $\hat{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{1}_n = 0$ for $i = 1, \dots, n_1 - 1$, we can conclude the result. ■

Proofs of Theorem 2 and Corollary 2. We write that $\boldsymbol{x}_j - \boldsymbol{\mu} = (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \sum_{m=1}^{k} \varepsilon_m(\boldsymbol{\mu}_i - \boldsymbol{\mu}_m)$ for $j = 1, \dots, n$; $i = 1, \dots, k$. We note that $\operatorname{Var}\{\boldsymbol{e}_d^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)/\Delta_{k-1}^{1/2}|\boldsymbol{x}_j \in \Pi_i\} = \boldsymbol{e}_d^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{e}_d/\Delta_{k-1} \le$

$\lambda_{\max}(\boldsymbol{\Sigma}_i)/\Delta_{k-1} = o(1)$ as $d \to \infty$ under Condition 1 for $j = 1, \dots, n; i = 1, \dots, k$, where $\boldsymbol{e}_d$ $(\in \mathbb{R}^d)$ is an arbitrary unit vector. Then, under Condition 1, when $\boldsymbol{x}_j \in \Pi_i$, it holds that

$$\frac{\boldsymbol{e}_d^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu})}{\Delta_{k-1}^{1/2}} = \frac{\boldsymbol{e}_d^{\mathrm{T}}\left\{\sum_{m=1}^{k}\varepsilon_m(\boldsymbol{\mu}_i - \boldsymbol{\mu}_m)\right\}}{\Delta_{k-1}^{1/2}} + o_P(1).$$

Then, using Lemmas 2 and 3, we can conclude the result of Theorem 2.

For the proof of Corollary 2, by noting that $\Delta_{i,i+1}/\Delta_i = 1 + o(1)$ and $\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\mu}_{i,i+1}/\Delta_{i,i+1}^{1/2} = \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\mu}_i/\Delta_i^{1/2} + o(1)$ for $i = 1, \dots, k-1$, under Condition 5, from Lemma 2, the results are obtained straightforwardly. ■

Proof of Theorem 3. By combining Lemmas 4 and 5, from Theorem 2 and the assumption that $\hat{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{z}_i \geq 0$ for all i, the result is obtained straightforwardly. ■

Proof of Proposition 2. Let $\boldsymbol{\Sigma}_{(*)} = \varepsilon_1\boldsymbol{\Sigma}_1 + \varepsilon_2\boldsymbol{\Sigma}_2$. Then, we define the eigendecomposition of $\boldsymbol{\Sigma}_{(*)}$ by $\boldsymbol{\Sigma}_{(*)} = \sum_{i=1}^{d}\lambda_{i(*)}\boldsymbol{h}_{i(*)}\boldsymbol{h}_{i(*)}^{\mathrm{T}}$, where $\lambda_{1(*)} \geq \cdots \geq \lambda_{d(*)} \geq 0$ are eigenvalues of $\boldsymbol{\Sigma}_{(*)}$ and $\boldsymbol{h}_{i(*)}$ is a unit eigenvector corresponding to $\lambda_{i(*)}$ for each i. Let $\lambda = \varepsilon_1\varepsilon_2\Delta_1$ and $\acute{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1/\Delta_1^{1/2}$. Then, from $\boldsymbol{\Sigma} = \lambda\acute{\boldsymbol{\mu}}_1\acute{\boldsymbol{\mu}}_1^{\mathrm{T}} + \boldsymbol{\Sigma}_{(*)}$, under $\max_{i=1,2}(\acute{\boldsymbol{\mu}}_1^{\mathrm{T}}\boldsymbol{\Sigma}_i\acute{\boldsymbol{\mu}}_1)/\Delta_1 \to 0$ as $d \to \infty$, it holds that $\acute{\boldsymbol{\mu}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\acute{\boldsymbol{\mu}}_1/\lambda \to 1$, so that

$$\sum_{i=1}^{d}\frac{\lambda_{i(*)}(\boldsymbol{h}_{i(*)}^{\mathrm{T}}\acute{\boldsymbol{\mu}}_1)^2}{\lambda} = o(1). \tag{C1}$$

Let $\kappa(i) = \lambda_{i(\star\star)} - \lambda$ for $i = 1, \dots, d$. For a sufficiently large d, when $\kappa(1) > 0$, there exists some positive integer $i_{\star\star}$ such that

$$i_{\star\star} = \max\{i|\kappa(i) > 0 \text{ for } i = 1, \dots, d\}.$$

Then, from Equation (C1), we have that $\sum_{i=1}^{i_{\star\star}}(\boldsymbol{h}_{i(*)}^{\mathrm{T}}\acute{\boldsymbol{\mu}}_1)^2 = o(1)$, so that $\lambda_{i_\star}/\lambda = 1 + o(1)$ with $i_\star = i_{\star\star} + 1$. When $\kappa(1) \leq 0$ for a sufficiently large d, it holds that $\lambda_{i_\star}/\lambda = 1 + o(1)$ with $i_\star = 1$. In addition, under $\liminf_{d\to\infty}|\lambda_{i'}/\lambda_{i_\star} - 1| > 0$ for $i' = 1, \dots, d(i' \neq i_\star)$, it holds that $\boldsymbol{h}_{i_\star}^{\mathrm{T}}\acute{\boldsymbol{\mu}}_1 = 1 + o(1)$ from $\boldsymbol{h}_{i_\star}^{\mathrm{T}}\boldsymbol{\mu}_1 \geq 0$. Then, from the fact that $\boldsymbol{h}_{i_\star}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{h}_{i_\star}/\lambda \to 0$ as $d \to \infty$ for $i = 1, 2$, in a way similar to Equation (A1), we have that $s_{i_\star j}/\lambda_{i_\star}^{1/2} = \boldsymbol{h}_{i_\star}^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu})/\lambda_{i_\star}^{1/2} = \boldsymbol{h}_{i_\star}^{\mathrm{T}}(\boldsymbol{\mu}_i - \boldsymbol{\mu})/\lambda_{i_\star}^{1/2} + o_P(1)$ when $\boldsymbol{x}_j \in \Pi_i$ for $j = 1, \dots, n; i = 1, 2$. We can conclude the results. ■