Journal of Multivariate Analysis 185 (2021) 104779

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings

Yugo Nakayama [a], Kazuyoshi Yata [b], Makoto Aoshima [b,*]

[a] Graduate School of Informatics, Kyoto University, Kyoto, Japan
[b] Institute of Mathematics, University of Tsukuba, Ibaraki, Japan

ARTICLE INFO

Article history:
Received 17 April 2020
Received in revised form 29 May 2021
Accepted 29 May 2021
Available online 8 June 2021

AMS 2000 subject classifications:
primary 62H25
secondary 62H30

Keywords:
HDLSS
Non-linear PCA
PC score
Radial basis function kernel
Spherical data

ABSTRACT

In this paper, we consider clustering based on the kernel principal component analysis (KPCA) for high-dimension, low-sample-size (HDLSS) data. We give theoretical reasons why the Gaussian kernel is effective for clustering high-dimensional data. In addition, we discuss a choice of the scale parameter yielding a high performance of the KPCA with the Gaussian kernel. Finally, we test the performance of the clustering by using microarray data sets.

1. Introduction

The clustering method is largely divided into hierarchical and non-hierarchical methods. For HDLSS data, Borysov et al. [10] and Kimes et al. [16] studied asymptotic behaviors of hierarchical clustering methods. Liu et al. [18], Ahn et al. [1], Huang et al. [14] and Sarkar and Ghosh [23] considered non-hierarchical type clustering methods. Especially, Liu et al. [18] proposed a binary split type clustering method called "statistical significance of clustering (SigClust)". On the other hand, principal component analysis (PCA) is a quite popular tool for non-hierarchical clustering of high dimensional data. Armstrong et al. [8] analyzed gene expression HDLSS data sets by clustering based on the linear PCA (LPCA). Yata and Aoshima [28] showed that the LPCA enjoys geometric consistency properties for the PC scores in high-dimensional mixture models. For non-linear data, the kernel PCA (KPCA) by Schölkopf et al. [24,25] is a non-linear extension of PCA by using kernel methods. There are a lot of data analyses based on the KPCA. For instance, Liu et al. [17] and Reverter et al. [22] analyzed HDLSS gene expression data by using the KPCA. However, as long as we know, asymptotic properties of the KPCA seem not to have been studied in HDLSS settings. In the current paper, we shall investigate asymptotic properties of the KPCA for HDLSS data.

Suppose there are independent and d-variate populations, Π_i, $i \in \{1, \ldots, k\}$, $k \geq 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ for each i. We assume $\limsup_{d\to\infty} \|\boldsymbol{\mu}_i\|^2/d < \infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i \in \{1, \ldots, k\}$, where $\|\cdot\|$ denotes the Euclidean norm. Here, for a function, $f(\cdot)$, "$f(d) \in (0, \infty)$ as $d \to \infty$" implies $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$. We mainly consider the case when $k = 2$. See Appendix C of the

* Correspondence to: Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan.
E-mail address: aoshima@math.tsukuba.ac.jp (M. Aoshima).

https://doi.org/10.1016/j.jmva.2021.104779

online supplementary for the case when $k = 3$. Suppose we have a $d \times n$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, where $\boldsymbol{x}_j$s are independently taken from Π_1 or Π_2. Let

$$n_i = \#\{j | \boldsymbol{x}_j \in \Pi_i \ \text{ for } j \in \{1, \ldots, n\}\},$$

where $\#A$ denotes the number of elements in a set A. Note that $n = n_1 + n_2$. We assume that n and n_is are independent of d, and $n_i \geq 1$ for $i \in \{1, 2\}$. For the sake of simplicity, we assume that $\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ and

$$\boldsymbol{x}_j \in \Pi_1,\ j \in \{1, \ldots, n_1\}, \quad \boldsymbol{x}_j \in \Pi_2,\ j \in \{n_1 + 1, \ldots, n\}. \tag{1}$$

In this paper, we study asymptotic properties of the kernel PCA in the HDLSS context that $d \to \infty$ while n is fixed. Let $\boldsymbol{K}$ be an $n \times n$ gram matrix with the (j, j') element $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \phi(\boldsymbol{x}_j)^\top \phi(\boldsymbol{x}_{j'})$, where $\phi(\cdot)$ is a feature map. Let $\boldsymbol{P}_n = \boldsymbol{I}_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^\top$, where $\boldsymbol{I}_n$ denotes the n-square identity matrix and $\boldsymbol{1}_n = (1, \ldots, 1)^\top$. We define the (centroid) gram matrix by

$$\boldsymbol{K}_0 = \boldsymbol{P}_n \boldsymbol{K} \boldsymbol{P}_n.$$

Note that $\mathrm{rank}(\boldsymbol{K}_0) \leq n - 1$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1}$ be the eigenvalues of $\boldsymbol{K}_0$. Then, we define the eigen-decomposition of $\boldsymbol{K}_0$ by

$$\boldsymbol{K}_0 = \sum_{i=1}^{n-1} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^\top,$$

where $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \ldots, \hat{u}_{in})^\top$ denotes a unit eigenvector corresponding to the $\hat{\lambda}_i$. Note that $\|\hat{\boldsymbol{u}}_i\| = 1$ and $\hat{\boldsymbol{u}}_i^\top \hat{\boldsymbol{u}}_{i'} = 0$ for all $i \neq i'$. The ith (normalized) PC score of $\boldsymbol{x}_j$ is given by $s_{ij} = \sqrt{n}\hat{u}_{ij}$. We note that $\sum_{j=1}^n s_{ij}^2/n = 1$ for all i. Also, note that $\sum_{j=1}^n s_{ij} = 0$ when $\hat{\lambda}_i > 0$ from the facts that $\boldsymbol{1}_n^\top \hat{\boldsymbol{u}}_i = \sum_{j=1}^n s_{ij}/\sqrt{n}$ and $\boldsymbol{1}_n^\top \boldsymbol{K}_0 \boldsymbol{1}_n = 0$. Since the sign of an eigenvector is arbitrary, we assume that $(\boldsymbol{1}_{n_1}^\top, -\boldsymbol{1}_{n_2}^\top)\hat{\boldsymbol{u}}_1 \geq 0$ without loss of generality.

We consider the following four typical kernels:

(I) The linear kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \boldsymbol{x}_j^\top \boldsymbol{x}_{j'}$;

(II) The Gaussian (radial basis function) kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma)$;

(III) The polynomial kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = (\zeta + \boldsymbol{x}_j^\top \boldsymbol{x}_{j'})^r$;

(IV) The Laplace kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|_1/\xi)$,

where $\gamma > 0$, $\zeta \geq 0$, $\xi > 0$, $r \in \mathbb{N}$ and $\|\cdot\|_1$ denotes the L_1-norm. Hellton and Thoresen [13], Shen et al. [26] and Yata and Aoshima [28] gave asymptotic properties of the PC score for the linear kernel function (I) in HDLSS settings. The Gaussian kernel function (II) is probably the most popular choice for kernel functions. Thus we mainly investigate the KPCA for the Gaussian kernel. In Appendix A, we give asymptotic properties of the KPCA in a general framework including the kernel functions (III) and (IV).

The rest of the paper is organized as follows: In Section 2, we provide motivations of the KPCA for HDLSS data. In Section 3, we give asymptotic properties of the KPCA with (I), that is, the LPCA. In Section 4, we give asymptotic properties of the KPCA with (II). We show that the KPCA gives better performances than the LPCA in HDLSS settings. In addition, we discuss a choice of the scale parameter γ yielding a high performance of the KPCA with (II). Finally, in Section 5, we examine the performance of the KPCA with (II) in numerical simulations and actual data analyses. All the theoretical results in this paper are given in the HDLSS context that $d \to \infty$ while n_is are fixed.

2. Motivations of the kernel PCA

In this section, let us give motivations of the KPCA for HDLSS data.

2.1. Kernel PCA for spherical data

We consider the following condition for $\boldsymbol{\Sigma}_i$, $i \in \{1, 2\}$,

$$\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2 = o(1),\ d \to \infty. \tag{2}$$

If we assume (2) and (A-i) given in Section 3, we have that as $d \to \infty$, when $\boldsymbol{x}_j \in \Pi_i$,

$$\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\| = \mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}\{1 + o_P(1)\}. \tag{3}$$

Thus, "$\boldsymbol{x}_j - \boldsymbol{\mu}_i$" concentrates on the surface of an expanding sphere with radius, $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$, as the dimension increases. See Dryden [11] and Hall et al. [12] for the details of the phenomenon. Aoshima and Yata [4,5] also proposed classifier based on the phenomenon. See also Aoshima et al. [3] for the review.

Remark 1. Let $\lambda_{i1} \geq \cdots \geq \lambda_{id}$ (≥ 0) be the eigenvalues of $\boldsymbol{\Sigma}_i$ $(i \in \{1, 2\})$. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2 = \sum_{\ell=1}^d \lambda_{i\ell}^2/(\sum_{\ell=1}^d \lambda_{i\ell})^2 \in [1/d, 1]$. Also, note that (2) is equivalent to "$\lambda_{i1}/\mathrm{tr}(\boldsymbol{\Sigma}_i) \to 0$ as $d \to \infty$". For instance, let us consider a spiked model as

$$\lambda_{i\ell} = a_{i\ell} d^{\alpha_{i\ell}},\ \ell \in \{1, \ldots, g\}, \quad \lambda_{i\ell} = c_{i\ell},\ \ell \in \{g + 1, \ldots, d\}, \tag{4}$$

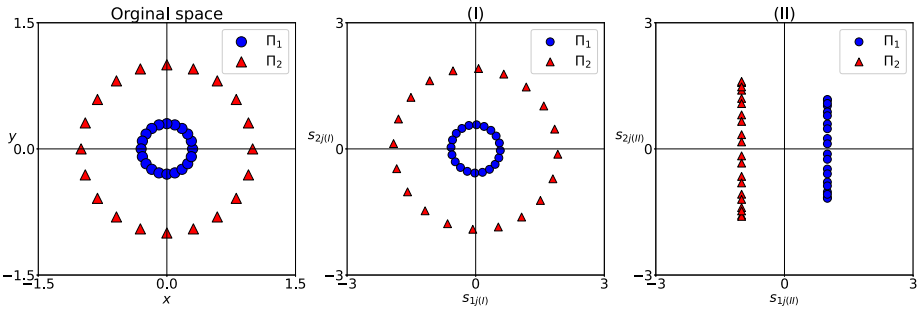


Fig. 1. Toy example to illustrate the kernel PC scores for spherical data. The left panel displays the original spherical data set, the center panel displays $(s_{1j(I)}, s_{2j(I)})$s and the right panel displays $(s_{1j(II)}, s_{2j(II)})$s.

where $a_{i\ell}$s, $c_{i\ell}$s and $\alpha_{i\ell}$s are positive (fixed) constants, and g is a positive (fixed) integer. See Yata and Aoshima [27] for the details of the spiked model. For (4), (2) holds when $\alpha_{i1} < 1$. On the other hand, (2) and (3) do not hold when $\alpha_{i1} = 1$. When $\alpha_{i1} = 1$, at least the first eigenvalue is strongly spiked and causes very huge noise. In this paper, we do not consider such cases as $\alpha_{i1} = 1$. Aoshima and Yata [6,7] developed a data transformation technique to avoid huge noise and to ensure high accuracy for inferences no matter how large the eigenvalues are. It would be possible to apply the data transformation technique to clustering for very noisy data and that is now under investigation.

Let us check the performance of the KPCA for spherical data. We considered the following toy example. Let $\theta_j = \pi(j/10)$ for $j \in \{1, \ldots, 20\}$. We generated 40 samples as $\boldsymbol{x}_j = (0.3\cos\theta_j, 0.3\sin\theta_j)^\top$ and $\boldsymbol{x}_{j+20} = (\cos\theta_j, \sin\theta_j)^\top$ for $j \in \{1, \ldots, 20\}$. Note that $\|\boldsymbol{x}_j\| = 0.3$ for $j \in \{1, \ldots, 20\}$, $\|\boldsymbol{x}_j\| = 1$ for $j \in \{21, \ldots, 40\}$, and $\sum_{j=1}^{20} \boldsymbol{x}_j = \sum_{j=21}^{40} \boldsymbol{x}_j = \boldsymbol{0}$. We calculated the first and second PC scores both for the linear kernel and Gaussian kernel with $\gamma = 1/2$. Let $s_{ij(I)}$ and $s_{ij(II)}$ denote s_{ij} given by using the kernel functions (I) and (II), respectively. In Fig. 1, we displayed scatter plots of $(s_{1j(I)}, s_{2j(I)})$ and $(s_{1j(II)}, s_{2j(II)})$, $j \in \{1, \ldots, 40\}$, together with scatter plots of the spherical data set itself. We observed that the linear kernel reproduces the spherical data set by $(s_{1j(I)}, s_{2j(I)})$s. On the other hand, the Gaussian kernel clustered the spherical data set by $s_{1j(II)}$s. It seems that the KPCA with (II) is useful for clustering spherical data.

2.2. Numerical behaviors of PC scores in HDLSS settings

Let

$$\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \quad \text{and} \quad \Delta_\Sigma = |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|.$$

Let us check the behavior of PC scores for several choices of Δ_μ, Δ_Σ and n_is. We considered the following toy examples. Independent pseudo random observations were generated from $\Pi_i : \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i \in \{1, 2\}$, with $\boldsymbol{\Sigma}_i = c_i \boldsymbol{I}_d$. We set $n = 20$, $\boldsymbol{\mu}_2 = \boldsymbol{0}$ and $c_1 = 1$. For $d = 100$, 1000 and 10 000, we considered four cases:

(a) $\boldsymbol{\mu}_1 = 3^{-1}\boldsymbol{1}_d$, $c_2 = 1$ and $(n_1, n_2) = (12, 8)$;
(b) $\boldsymbol{\mu}_1 = 3^{-1}\boldsymbol{1}_d$, $c_2 = 2$ and $(n_1, n_2) = (12, 8)$;
(c) $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $c_2 = 2$ and $(n_1, n_2) = (12, 8)$;
(d) $\boldsymbol{\mu}_1 = 3^{-1}\boldsymbol{1}_d$, $c_2 = 2$ and $(n_1, n_2) = (8, 12)$.

Note that $\Delta_\mu = d/9$ for (a), (b) and (d), and $\Delta_\mu = 0$ for (c). Also, note that $\Delta_\Sigma = d$ for (b) to (d) and $\Delta_\Sigma = 0$ for (a). In Figs. 2–3, we displayed scatter plots of $(s_{1j(I)}, s_{2j(I)})$ and $(s_{1j(II)}, s_{2j(II)})$, $j \in \{1, \ldots, n\}$, together with two vertical lines, $\sqrt{n_2/n_1}$ and $-\sqrt{n_1/n_2}$. See Sections 3 and 4 for the details of the lines. We observed that $s_{1j(II)}$ became close to $(-1)^{i+1}\sqrt{n_{i'}/n_i}$ $(i' \neq i)$ when $\boldsymbol{x}_j \in \Pi_i$ for all j as d increases. In addition, $s_{2j(II)}$ became close to 0 when $\boldsymbol{x}_j \in \Pi_1$ as d increases for (b) to (d). On the other hand, for (c) and (d), $s_{ij(I)}$s behaved totally different from $s_{ij(II)}$s. We shall explain their theoretical backgrounds in Sections 3 and 4.

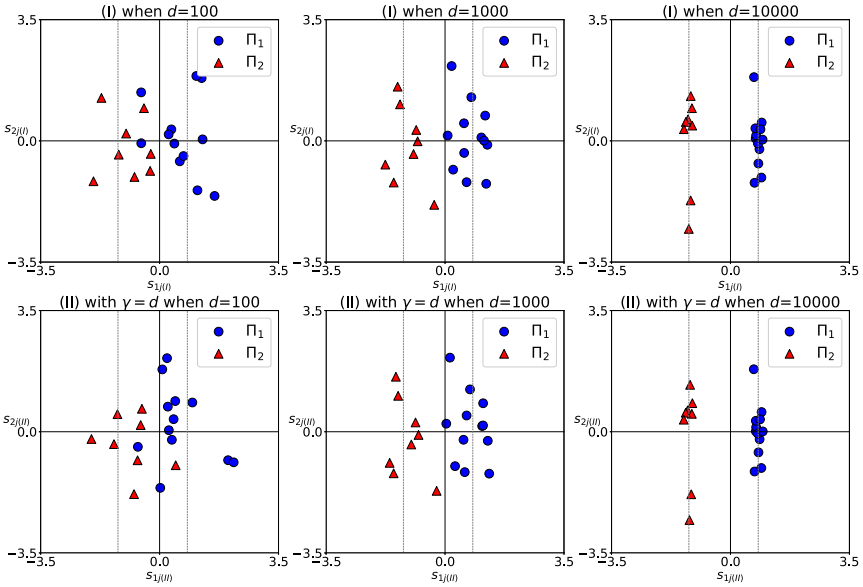
3. Kernel PC with the linear kernel (I)

In this section, we consider the KPCA with (I), that is, the LPCA. We assume the following condition:

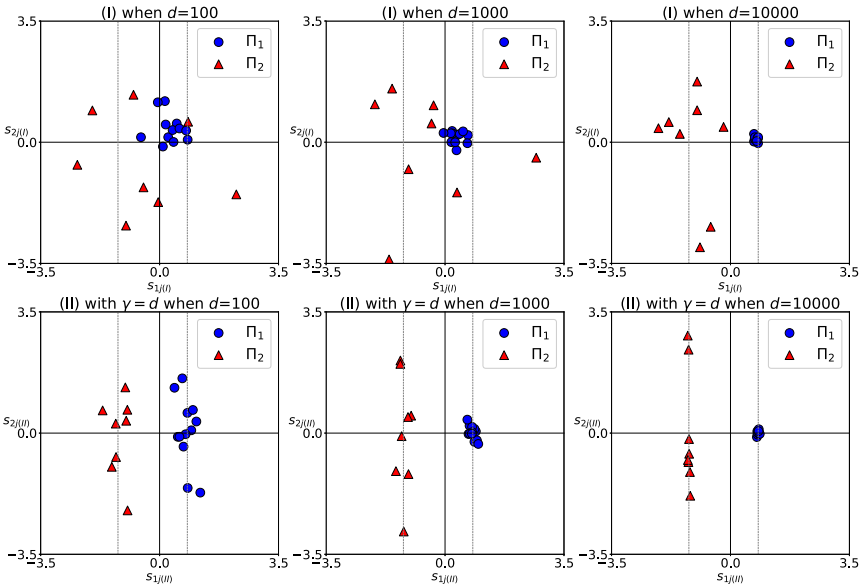
(A-i) $\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$ as $d \to \infty$ for $i \in \{1, 2\}$.

Note that $\mathrm{E}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i) = \mathrm{tr}(\boldsymbol{\Sigma}_i)$ for $i \in \{1, 2\}$. If Π_is are Gaussian, it holds that $\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i) = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ for $i \in \{1, 2\}$, so that (A-i) naturally holds.

Remark 2. We denote the eigen-decomposition of $\boldsymbol{\Sigma}_i$ $(i \in \{1, 2\})$ by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^\top$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{id})$ having eigenvalues, $\lambda_{i1} \geq \cdots \geq \lambda_{id} \geq 0$, and $\boldsymbol{H}_i$ is an orthogonal matrix of the corresponding eigenvectors. When $\boldsymbol{x} \in \Pi_i$ $(i \in$



(a) When $\boldsymbol{\mu}_1 = 3^{-1}\mathbf{1}_d$, $c_2 = 1$ and $(n_1, n_2) = (12, 8)$.



(b) When $\boldsymbol{\mu}_1 = 3^{-1}\mathbf{1}_d$, $c_2 = 2$ and $(n_1, n_2) = (12, 8)$.

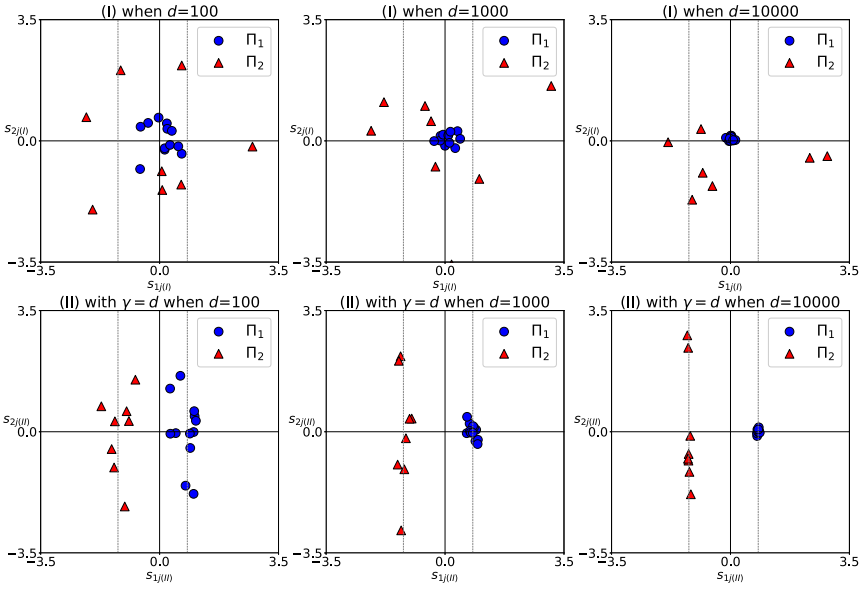
Fig. 2. Toy example to illustrate the behaviors of the PC scores by the linear and Gaussian kernels for (a) and (b). The blue circles and red triangles denote the data points belonging to Π_1 and Π_2, respectively. The upper (lower) three panels illustrate the linear (Gaussian) kernel for (a) and (b).

$\{1, 2\}$), let us write that $\boldsymbol{x} - \boldsymbol{\mu}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2}(z_{i1}, \ldots, z_{id})^\top$. Note that $\mathrm{E}\{(z_{i1}, \ldots, z_{id})^\top\} = \mathbf{0}$ and $\mathrm{Var}\{(z_{i1}, \ldots, z_{id})^\top\} = \boldsymbol{I}_d$. Then, it holds that

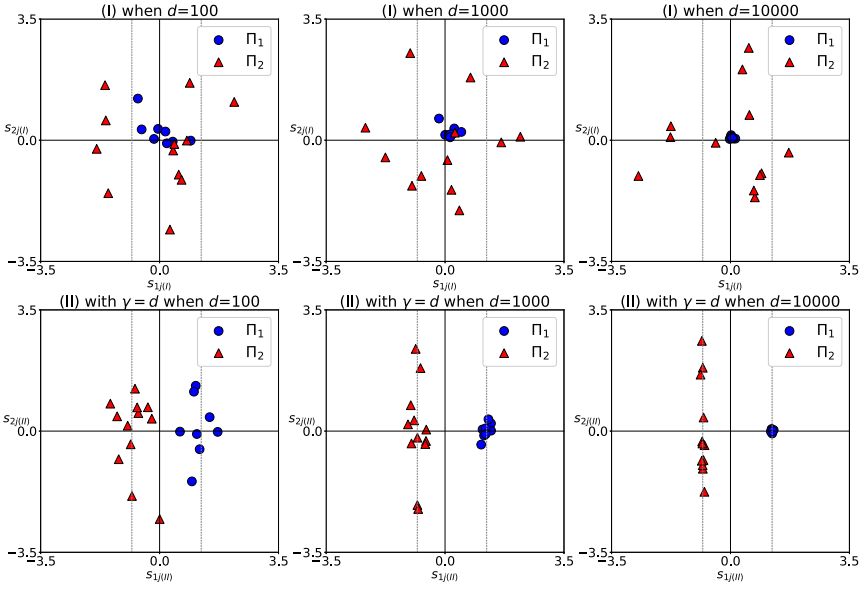
$$\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i) = \sum_{\ell, \ell'=1}^{d} \lambda_{i\ell}\lambda_{i\ell'}\mathrm{E}\{(z_{i\ell}^2 - 1)(z_{i\ell'}^2 - 1)\}.$$

If $\limsup_{d\to\infty} \mathrm{E}(z_{i\ell}^4) < \infty$ and $\mathrm{E}(z_{i\ell}^2 z_{i\ell'}^2) = \mathrm{E}(z_{i\ell}^2)\mathrm{E}(z_{i\ell'}^2)$ for all $\ell \neq \ell'$, (A-i) holds. Thus if Π_is are Gaussian, (A-i) holds since $z_{i\ell}$s are independent and identically distributed (i.i.d.) as the standard normal distribution when Π_i is Gaussian.

We write that $\boldsymbol{K}_{0(I)} = \boldsymbol{P}_n \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{P}_n = (\boldsymbol{X} - \overline{\boldsymbol{X}})^\top (\boldsymbol{X} - \overline{\boldsymbol{X}})$, where $\overline{\boldsymbol{X}} = (\overline{\boldsymbol{x}}, \ldots, \overline{\boldsymbol{x}})$ with $\overline{\boldsymbol{x}} = \sum_{j=1}^n \boldsymbol{x}_j/n$. We have the sample covariance matrix as $\boldsymbol{S} = (\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^\top/(n-1)$. Then, its dual covariance matrix, $\boldsymbol{K}_{0(I)}/(n-1)$, shares non-zero eigenvalues with $\boldsymbol{S}$. See Ahn et al. [2], Jung and Marron [15] and Yata and Aoshima [27] for the details of the dual



(c) When $\boldsymbol{\mu}_1 = \mathbf{0}$, $c_2 = 2$ and $(n_1, n_2) = (12, 8)$.



(d) When $\boldsymbol{\mu}_1 = 3^{-1}\mathbf{1}_d$, $c_2 = 2$ and $(n_1, n_2) = (8, 12)$.

Fig. 3. Toy example to illustrate the behaviors of the PC scores by the linear and Gaussian kernels for (c) and (d). The blue circles and red triangles denote the data points belonging to Π_1 and Π_2, respectively. The upper (lower) three panels illustrate the linear (Gaussian) kernel for (c) and (d).

covariance matrix. We define the eigen-decomposition of $\boldsymbol{K}_{0(I)}$ by $\boldsymbol{K}_{0(I)} = \sum_{i=1}^{n-1} \hat{\lambda}_{i(I)} \hat{\boldsymbol{u}}_{i(I)} \hat{\boldsymbol{u}}_{i(I)}^\top$, where $\hat{\lambda}_{1(I)} \geq \cdots \geq \hat{\lambda}_{n-1(I)}$ and $\hat{\boldsymbol{u}}_{i(I)} = (\hat{u}_{i1(I)}, \ldots, \hat{u}_{in(I)})^\top$. Note that $s_{ij(I)} = \sqrt{n}\hat{u}_{ij(I)}$. Let $\boldsymbol{r} = n^{-1}(n_2 \mathbf{1}_{n_1}^\top, -n_1 \mathbf{1}_{n_2}^\top)^\top$ and

$$\widetilde{\boldsymbol{K}}_{0(I)} = \Delta_\mu \boldsymbol{r}\boldsymbol{r}^\top + \boldsymbol{P}_n \begin{pmatrix} \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_{n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & \mathrm{tr}(\boldsymbol{\Sigma}_2)\boldsymbol{I}_{n_2} \end{pmatrix} \boldsymbol{P}_n,$$

where $\boldsymbol{O}_{n_1,n_2}$ denotes the $n_1 \times n_2$ zero matrix. Let $\|\cdot\|_F$ denote the Frobenius norm. Then, we have the following result.

Lemma 1. *Assume (A-i). Assume also*

(A-ii) $tr(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2 = o(1)$ *as* $d \to \infty$ *for* $i \in \{1, 2\}$.

Then, it holds that $\|\boldsymbol{K}_{0(I)} - \widetilde{\boldsymbol{K}}_{0(I)}\|_F = o_P(\Delta_\mu)$ *as* $d \to \infty$.

Here, "$\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2 = o(1)$" means that the intrinsic information about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is much larger than the noise. If $\liminf_{d\to\infty} \Delta_\mu/d > 0$, (A-ii) is equivalent to (2).

Now, we consider the eigenstructure of $\widetilde{\boldsymbol{K}}_{0(I)}$. Note that

$$\begin{pmatrix} \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_{n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & \mathrm{tr}(\boldsymbol{\Sigma}_2)\boldsymbol{I}_{n_2} \end{pmatrix} = \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_n + \Delta_\Sigma \begin{pmatrix} \boldsymbol{O}_{n_1,n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & \boldsymbol{I}_{n_2} \end{pmatrix}$$

from (1). By noting that $\boldsymbol{P}_n\boldsymbol{r} = \boldsymbol{r}$, when $n_2 \geq 2$, we write that

$$\boldsymbol{P}_n \begin{pmatrix} \boldsymbol{O}_{n_1,n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & \boldsymbol{I}_{n_2} \end{pmatrix} \boldsymbol{P}_n = \sum_{i=1}^{n_2-1} \boldsymbol{v}_i\boldsymbol{v}_i^\top + \frac{n_1}{n}\frac{\boldsymbol{r}\boldsymbol{r}^\top}{\|\boldsymbol{r}\|^2}, \tag{5}$$

where the set of vectors, $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_2-1}, \boldsymbol{r}/\|\boldsymbol{r}\|\}$, is orthonormal. Then, by noting that $\|\boldsymbol{r}\|^2 = n_1n_2/n$, we write that

$$\widetilde{\boldsymbol{K}}_{0(I)} = (n_1n_2\Delta_\mu/n + n_1\Delta_\Sigma/n)\frac{\boldsymbol{r}\boldsymbol{r}^\top}{\|\boldsymbol{r}\|^2} + \Delta_\Sigma \sum_{i=1}^{n_2-1} \boldsymbol{v}_i\boldsymbol{v}_i^\top + \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{P}_n. \tag{6}$$

Note that $n_1n_2\Delta_\mu/n + n_1\Delta_\Sigma/n + \mathrm{tr}(\boldsymbol{\Sigma}_1)$ and $\Delta_\Sigma + \mathrm{tr}(\boldsymbol{\Sigma}_1)$ are eigenvalues of $\widetilde{\boldsymbol{K}}_{0(I)}$ when $n_2 \geq 2$. Also, note that

$$\frac{n_1n_2\Delta_\mu/n + n_1\Delta_\Sigma/n - \Delta_\Sigma}{\Delta_\mu} = \frac{n_1n_2}{n}\left(1 - \frac{\Delta_\Sigma}{n_1\Delta_\mu}\right). \tag{7}$$

Here, we assume the following condition:

(A-iii) $\limsup_{d\to\infty} \dfrac{\Delta_\Sigma}{n_1\Delta_\mu} < 1$ when $n_2 \geq 2$.

From (6) and (7), under (A-iii), the first unit eigenvector of $\widetilde{\boldsymbol{K}}_{0(I)}$ is $\boldsymbol{r}/\|\boldsymbol{r}\|$. Thus from Lemma 1, under (A-i) and (A-ii), $\hat{\boldsymbol{u}}_{1(I)}$ tends to $\boldsymbol{r}/\|\boldsymbol{r}\| = (n_2\boldsymbol{1}_{n_1}^\top, -n_1\boldsymbol{1}_{n_2}^\top)^\top/\sqrt{n_1n_2n}$ as $d \to \infty$. Then, we have the following result.

Theorem 1. *Assume (A-i) to (A-iii). Then, it holds that as* $d \to \infty$

$$s_{1j(I)} = \begin{cases} \sqrt{n_2/n_1} + o_P(1), & j \in \{1, \ldots, n_1\}, \\ -\sqrt{n_1/n_2} + o_P(1), & j \in \{n_1+1, \ldots, n\}. \end{cases} \tag{8}$$

Remark 3. Yata and Aoshima [28] gave the results similar to Theorem 1 under $\Delta_\Sigma/\Delta_\mu = o(1)$ as $d \to \infty$. Note that (A-iii) is milder than $\Delta_\Sigma/\Delta_\mu = o(1)$. When $n_2 = 1$, (8) holds under (A-i) and (A-ii). From Corollary 3 in [27] and Eq. (3) in [28], under (A-ii) and some regularity conditions, it holds that

$$\mathrm{Angle}(\hat{\boldsymbol{u}}_{1(I)}, \boldsymbol{r}) = o_P(1) \text{ as } d \to \infty \text{ and } n \to \infty. \tag{9}$$

Thus, the data can be effectively classified by the sign of the first PC scores even when $d \to \infty$ and $n \to \infty$.

Remark 4. If $\mathrm{tr}(\boldsymbol{\Sigma}_1) > \mathrm{tr}(\boldsymbol{\Sigma}_2)$, (8) holds as $d \to \infty$ under (A-i), (A-ii) and the following condition:

$$\limsup_{d\to\infty} \Delta_\Sigma/(n_2\Delta_\mu) < 1 \quad \text{when } n_1 \geq 2.$$

By using Theorem 1, one can classify $\boldsymbol{x}_j$s into two groups by the sign of the first PC scores. Note that $\Delta_\Sigma/(n_2\Delta_\mu) = 0$ and $\Delta_\Sigma/(n_1\Delta_\mu) = 3/4 < 1$ in the settings (a) and (b) of Fig. 2, respectively. Thus from Theorem 1, as expected theoretically, $s_{1j(I)}$ became close to $(-1)^{i+1}\sqrt{n_{i'}/n_i}$ $(i' \neq i)$ when $\boldsymbol{x}_j \in \Pi_i$ for all j, in Fig. 2.

In addition, we have the following result for the i (≥ 2)th PC score.

Proposition 1. *Assume (A-i) to (A-iii). Assume also* $n_2 \geq 2$ *and*

(A-iv) $\liminf_{d\to\infty} \dfrac{\Delta_\Sigma}{\Delta_\mu} > 0$.

Then, it holds that as $d \to \infty$

$$\sum_{j'=n_1+1}^{n} \frac{s_{ij'(I)}^2}{n} = 1 + o_P(1), \ s_{ij(I)} = o_P(1), \ j \in \{1, \ldots, n_1\}, \ i \in \{2, \ldots, n_2\}.$$

We note that under the assumptions in Proposition 1, $\hat{\boldsymbol{u}}_{i(I)}$ tends to a linear combination of $\boldsymbol{v}_{i'}$, $i' \in \{1, \ldots, n_2 - 1\}$, as $d \to \infty$ for $i \in \{2, \ldots, n_2\}$. From Proposition 1, $s_{2j(I)}$ concentrates on 0 for $\boldsymbol{x}_j \in \Pi_1$. Thus, under the assumptions in Proposition 1, one can classify $\boldsymbol{x}_j$s into two groups, effectively, by the first two PC scores. See the upper three panels of (b) in Fig. 2. On the other hand, if (A-iii) is not met, we have the following result.

Proposition 2. *Assume (A-i) and* $tr(\boldsymbol{\Sigma}_i^2)/\Delta_\Sigma^2 = o(1)$ *as* $d \to \infty$ *for* $i = 1, 2$. *Assume also* $n_2 \geq 2$ *and*

$$\liminf_{d\to\infty} \frac{\Delta_\Sigma}{n_1 \Delta_\mu} > 1.$$

Then, it holds that as $d \to \infty$

$$\sum_{j'=n_1+1}^{n} \frac{s_{ij'(I)}^2}{n} = 1 + o_P(1),\ s_{ij(I)} = o_P(1),\ j \in \{1, \ldots, n_1\}, i \in \{1, \ldots, n_2 - 1\}.$$

Remark 5. Yata and Aoshima [28] gave the results similar to Proposition 2 under $\Delta_\mu/\Delta_\Sigma = o(1)$ as $d \to \infty$.

We note that under the assumptions in Proposition 2, $\hat{\boldsymbol{u}}_{i(I)}$ tends to a linear combination of $\boldsymbol{v}_{i'}$, $i' \in \{1, \ldots, n_2 - 1\}$, as $d \to \infty$ for $i \in \{1, \ldots, n_2 - 1\}$. Also, note that (A-iii) is not met for the settings (c) and (d) of Fig. 3 since $\Delta_\Sigma/(n_1\Delta_\mu) = 9/8 > 1$ for (d). Thus, as expected theoretically, $s_{ij(I)}$, $i \in \{1, 2\}$, became close to 0 for $\boldsymbol{x}_j \in \Pi_1$ in Fig. 3. However, it is difficult to cluster $\boldsymbol{x}_j$s into two groups in such cases. Therefore, if Δ_μ/Δ_Σ is small, we do not recommend to use the linear PCA.

4. Kernel PCA with the Gaussian kernel (II)

In this section, we consider the KPCA with (II).

4.1. Asymptotic properties of the PC scores

Let $\boldsymbol{K}_{0(II)}$ denote $\boldsymbol{K}_0$ given by using the kernel function (II). We assume the following condition for γ in (II):

(A-v) $\limsup_{d\to\infty} \dfrac{\Delta_\mu + \Delta_\Sigma}{\gamma} < \infty.$

Note that (A-v) holds under the following condition:

$$\limsup_{d\to\infty} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2)}{\gamma} < \infty. \tag{10}$$

See also Remark 6 for the details of (A-v). Let $\kappa_i = \exp\{-\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma\}$ for $i \in \{1, 2\}$, $\kappa_\mu = \exp(-\Delta_\mu/\gamma)$ and $\kappa_\Sigma = \exp(-\Delta_\Sigma/\gamma)$. Note that $\kappa_1 \geq \kappa_2$ from $\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$. Also, note that $\kappa_\mu < 1$ and $\kappa_\Sigma < 1$ when $\Delta_\mu \neq 0$ and $\Delta_\Sigma \neq 0$, respectively. Let $\Delta_\kappa = 1 + \kappa_\Sigma^2 - 2\kappa_\mu\kappa_\Sigma$ and

$$\widetilde{\boldsymbol{K}}_{0(II)} = \Delta_\kappa \kappa_1^2 \boldsymbol{r}\boldsymbol{r}^\top + \boldsymbol{P}_n \begin{pmatrix} (1-\kappa_1^2)\boldsymbol{I}_{n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & (1-\kappa_1^2\kappa_\Sigma^2)\boldsymbol{I}_{n_2} \end{pmatrix} \boldsymbol{P}_n.$$

Here, Δ_κ is a distance between the two populations since $\Delta_\kappa = (1-\kappa_\Sigma)^2 + 2(1-\kappa_\mu)\kappa_\Sigma \geq 0$, and $\Delta_\kappa > 0$ when $\Delta_\mu \neq 0$ or $\Delta_\Sigma \neq 0$. Then, we have the following result.

Lemma 2. *Assume (A-i) and (A-v). Assume also*

(A-vi) $\dfrac{tr(\boldsymbol{\Sigma}_i^2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\gamma^2 \Delta_\kappa^2} = o(1),\ d \to \infty,\ i \in \{1, 2\}.$

Then, it holds that $\|\boldsymbol{K}_{0(II)} - \widetilde{\boldsymbol{K}}_{0(II)}\|_F = o_P(\Delta_\kappa \kappa_1^2)$, $d \to \infty$.

Remark 6. We note that (A-vi) is a convergence condition of $\boldsymbol{K}_{0(II)}$. Note that $\liminf_{d\to\infty} \kappa_\mu > 0$ and $\liminf_{d\to\infty} \kappa_\Sigma > 0$ under (A-v). Then, it holds that $\liminf_{d\to\infty} \gamma\Delta_\kappa/\Delta_\mu > 0$ under (A-v), so that (A-vi) is a milder condition than (A-ii) under (A-v) from the fact that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leq \Delta_\mu \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$. However, (A-vi) is not always milder than (A-ii) when (A-v) is not met. In addition, when (A-v) is not met, the KPCA with (II) gave bad performances in Figs. 4–5 in Section 5. Thus, we assume (A-v) for γ in (II).

From Remark 6, (A-vi) holds under (A-ii) and (A-v). We have the following result.

Proposition 3. *(A-vi) holds under (A-v) and the condition:*

$$\frac{tr(\boldsymbol{\Sigma}_i^2)}{\max\{\Delta_\Sigma^2/\gamma, \Delta_\mu\}^2} = o(1),\ d \to \infty,\ i \in \{1, 2\}. \tag{11}$$

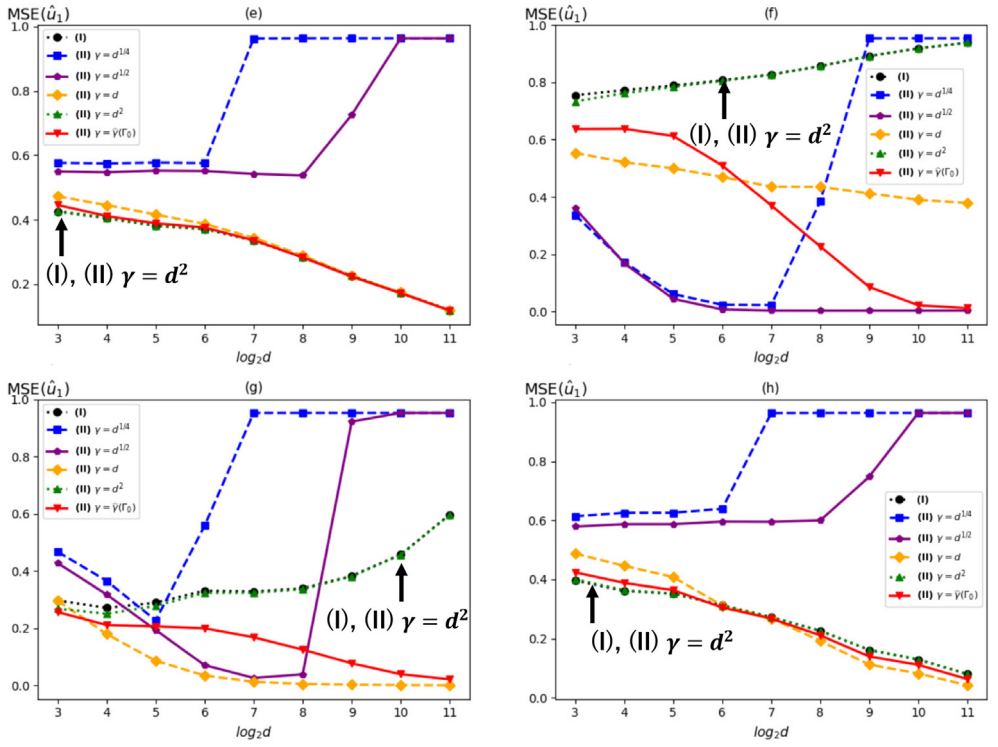


Fig. 4. The performances of the PC scores with the kernel functions (I) and (II) for (e) to (h). For the Gaussian kernel, we set $\gamma = d^{2^s/8}$ ($s \in \{1, \ldots, 4\}$) and $\hat{\gamma}(\Gamma_0)$.

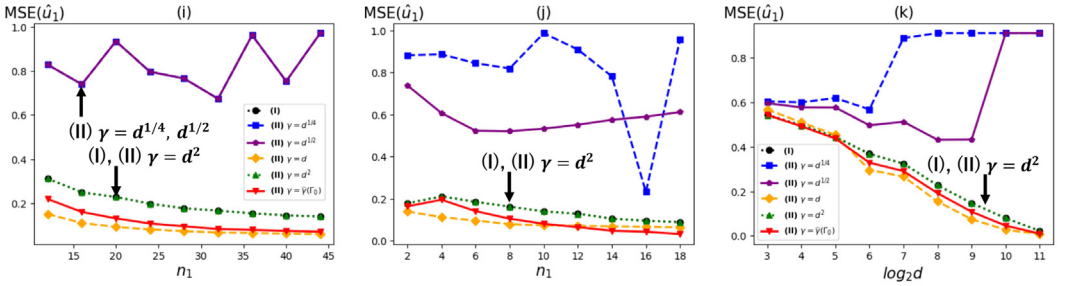


Fig. 5. The performances of the PC scores with the kernel functions (I) and (II) for (i) to (k). For the Gaussian kernel, we set $\gamma = d^{2^s/8}$ ($s \in \{1, \ldots, 4\}$) and $\hat{\gamma}(\Gamma_0)$.

Note that (A-vi) holds even when $\Delta_\mu = 0$. From (5), when $n_2 \geq 2$, we write that

$$\widetilde{\boldsymbol{K}}_{0(II)} = \kappa_1^2\big(n_1 n_2 \Delta_\kappa / n + n_1(1-\kappa_\Sigma^2)/n\big)\frac{\boldsymbol{r}\boldsymbol{r}^\top}{\|\boldsymbol{r}\|^2} + \kappa_1^2(1-\kappa_\Sigma^2)\sum_{i=1}^{n_2-1}\boldsymbol{v}_i\boldsymbol{v}_i^\top + (1-\kappa_1^2)\boldsymbol{P}_n. \tag{12}$$

Similar to (7), from Lemma 2, for the PC score by the kernel function (II), we have the following result.

Theorem 2. *Assume (A-i), (A-v) and (A-vi). Assume also*

(A-vii) $\limsup_{d\to\infty} \dfrac{(1-\kappa_\Sigma^2)}{n_1\Delta_\kappa} < 1$ *when* $n_2 \geq 2$.

Then, it holds as $d \to \infty$

$$s_{1j(II)} = \begin{cases} \sqrt{n_2/n_1} + o_P(1), & j \in \{1, \ldots, n_1\}, \\ -\sqrt{n_1/n_2} + o_P(1), & j \in \{n_1+1, \ldots, n\}. \end{cases} \tag{13}$$

Furthermore, it holds as $d \to \infty$

$$\sum_{j'=n_1+1}^{n} \frac{s^2_{ij'(II)}}{n} = 1 + o_P(1),\ s_{ij(II)} = o_P(1),\ j \in \{1, \ldots, n_1\},\ i \in \{2, \ldots, n_2\}$$

under $n_2 \geq 2$ *and the condition:*

(A-viii) $\liminf_{d\to\infty} \dfrac{(1-\kappa_\Sigma^2)}{\Delta_\kappa} > 0.$

Remark 7. Similar to (9), for the KPCA with (II) it would be possible to obtain "Angle$(\hat{\boldsymbol{u}}_1, \boldsymbol{r}) = o_p(1)$ as $d \to \infty$ and $n \to \infty$" and that is now under investigation. We provide some simulation studies for large n_is in Fig. 5.

From Theorem 2 and Proposition 3, the PC score with (II) has the consistency (13) even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Note that (A-vii) holds for (a) to (d) in Figs. 2–3. Thus in Figs. 2–3, $s_{1j(II)}$ became close to $(-1)^{i+1}\sqrt{n_{i'}/n_i}$ $(i' \neq i)$ as d increases when $\boldsymbol{x}_j \in \Pi_i$. On the other hand, the PC score with (I) does not hold the consistency property when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. See (c) in Fig. 3.

When (A-vii) is not met, we have the following result.

Proposition 4. *Assume (A-i), (A-v) and*

$$\frac{tr(\boldsymbol{\Sigma}_i^2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\gamma^2(1-\kappa_\Sigma^2)^2} = o(1),\ d \to \infty,\ i \in \{1, 2\}.$$

Assume also $n_2 \geq 2$ *and*

$$\liminf_{d\to\infty} \frac{(1-\kappa_\Sigma^2)}{n_1\Delta_\kappa} > 1.$$

Then, it holds that as $d \to \infty$

$$\sum_{j'=n_1+1}^{n} \frac{s^2_{ij'(II)}}{n} = 1 + o_P(1),\ s_{ij(II)} = o_P(1),\ j \in \{1, \ldots, n_1\},\ i \in \{1, \ldots, n_2 - 1\}.$$

From Theorem 2, one can classify $\boldsymbol{x}_j$s into two groups by the PC score under (A-vii) and the regularity conditions. Thus we recommend to use γ satisfying (A-vii).

4.2. Relation between the linear kernel and Gaussian kernel

For $\widetilde{\boldsymbol{K}}_{0(I)}$ and $\widetilde{\boldsymbol{K}}_{0(II)}$, we have the following result.

Proposition 5. *Under* $\max_{i=1,2} tr(\boldsymbol{\Sigma}_i)/\gamma = o(1)$ *and* $\Delta_\Sigma^2/(\gamma\Delta_\mu) = o(1)$ *as* $d \to \infty$, *it holds that as* $d \to \infty$

$$\|\widetilde{\boldsymbol{K}}_{0(I)}/\Delta_\mu - \widetilde{\boldsymbol{K}}_{0(II)}/(\kappa_1^2\Delta_\kappa)\|_F = o(1).$$

From Proposition 5, under $\max_{i=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma = o(1)$ and $\Delta_\Sigma^2/(\gamma\Delta_\mu) = o(1)$ as $d \to \infty$, the PC score with the Gaussian kernel function (II) is asymptotically equivalent to that with the linear kernel function (I).

4.3. How to choose γ

In this section, we discuss a choice of γ in the Gaussian kernel function (II). Let $\eta_i = n_i/n$ for $i \in \{1, 2\}$. We assume $n_2 \geq 3$ and $\Delta_\kappa > 0$ in this section. Let $\lambda_{1(II)} \geq \cdots \geq \lambda_{n-1(II)}$ be the eigenvalues of $\widetilde{\boldsymbol{K}}_{0(II)}$. Let $\alpha_1 = (1-\kappa_1^2)/(\Delta_\kappa\kappa_1^2)$, $\alpha_2 = (1-\kappa_\Sigma^2)/\Delta_\kappa$ and $\beta_j = \lambda_{j(II)}/(\Delta_\kappa\kappa_1^2)$ for $j \in \{1, \ldots, n-1\}$. From (12) we have that

$$\beta_1 = \alpha_1 + \eta_1 n_2 + \eta_1\alpha_2,\quad \beta_2 = \cdots = \beta_{n_2} = \alpha_1 + \alpha_2,\quad \text{when } \alpha_2/n_1 < 1;$$
$$\text{and}\quad \beta_1 = \cdots = \beta_{n_2-1} = \alpha_1 + \alpha_2,\quad \beta_{n_2} = \alpha_1 + \eta_1 n_2 + \eta_1\alpha_2,\quad \text{when } \alpha_2/n_1 \geq 1.$$

From the above results, if $\liminf_{d\to\infty}(\beta_1 - \beta_2) > 0$, (A-vii) holds. On the other hand, if $\beta_1 = \beta_2$, (A-vii) does not hold. Also, from Theorem 2, we emphasize that the first eigenspace includes the intrinsic information for the two-class model. If the difference between β_1 and β_2 is large, the intrinsic information becomes clearer. Thus one may choose γ that makes "$\lambda_{1(II)}(\gamma) - \lambda_{2(II)}(\gamma)$" large, where $\lambda_{j(II)}(\gamma)$ denotes $\lambda_{j(II)}$ for a given tuning parameter γ. We propose the following procedure to choose γ: We denote a set of candidates of γ by $\Gamma = \{\gamma_1, \ldots, \gamma_t\}$. Let $\hat{\lambda}_{1(II)}(\gamma) \geq \cdots \geq \hat{\lambda}_{n-1(II)}(\gamma)$ be the eigenvalues of $\boldsymbol{K}_{0(II)}$ with a given γ. Choose γ such that

$$\hat{\gamma}(\Gamma) = \underset{\gamma\in\Gamma}{\mathrm{argmax}}(\hat{\lambda}_{1(II)}(\gamma) - \hat{\lambda}_{2(II)}(\gamma)). \tag{14}$$

We give the performance of $\hat{\gamma}(\Gamma)$ in Section 5. Let $\gamma_* = \mathrm{tr}(\boldsymbol{S})$. From (10), γ_* is a candidate of γ since $\mathrm{E}\{\mathrm{tr}(\boldsymbol{S})\} = \eta_1\mathrm{tr}(\boldsymbol{\Sigma}_1) + \eta_2\mathrm{tr}(\boldsymbol{\Sigma}_2) + \eta_1\eta_2\Delta_\mu$. We recommend to consider γs around γ_* as candidates in Γ. See (15) in Section 5.1. We provide in Appendix D of the online supplementary material a program in R-code to calculate $s_{ij(II)}$s with $\gamma = \hat{\gamma}(\Gamma)$.

5. Performances

In this section, we check the performance of the KPCA both in numerical simulations and actual data analyses.

5.1. Numerical studies

For the toy examples in Figs. 2 to 3 of Section 2.2, we checked the performance of $s_{ij(II)}$s with $\gamma = d$. In this section, we checked the performance of $s_{ij(II)}$s with $\gamma = \hat{\gamma}(\Gamma)$ in the following setup. We set Γ in (14) as

$$\Gamma_0 = \left\{\gamma_*^{t/5},\ t \in \{1, \ldots, 9\}\right\}. \tag{15}$$

Independent pseudo random observations were generated from $\Pi_i : \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, having $\boldsymbol{\Sigma}_1 = c_1\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = c_2\boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B} = \mathrm{diag}[\{0.5 + 1/(d+1)\}^{1/2}, \ldots, \{0.5 + d/(d+1)\}^{1/2}]$. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and $d = 2^s,\ s = 3, \ldots, 11$. Let $\boldsymbol{\mu}_* = (1, \ldots, 1, 0, \ldots, 0)^\top$ whose first $\lceil d^{2/3} \rceil$ elements are 1. Here, $\lceil \cdot \rceil$ denotes the ceiling function. We considered four cases:

(e) $n_1 = 12,\ n_2 = 3,\ \boldsymbol{\mu}_2 = \boldsymbol{\mu}_*$ and $(c_1, c_2) = (1, 1)$;
(f) $n_1 = 3,\ n_2 = 12,\ \boldsymbol{\mu}_2 = \boldsymbol{0}$ and $(c_1, c_2) = (0.3, 1)$;
(g) $n_1 = 3,\ n_2 = 12,\ \boldsymbol{\mu}_2 = 2\boldsymbol{\mu}_*$ and $(c_1, c_2) = (1, 2)$;
(h) $n_1 = 12,\ n_2 = 3,\ \boldsymbol{\mu}_2 = \boldsymbol{\mu}_*$ and $(c_1, c_2) = (1, 1 + 2/\lceil d^{1/3} \rceil)$.

Note that $\Delta_\mu \approx d^{2/3}$ for (e) and (h), $\Delta_\mu = 0$ for (f) and $\Delta_\mu \approx 4d^{2/3}$ for (g). Also, note that $\Delta_\Sigma = 0$ for (e), $\Delta_\Sigma = 0.7d$ for (f), $\Delta_\Sigma = d$ for (g) and $\Delta_\Sigma \approx 2d^{2/3}$ for (h). We calculated $\hat{\boldsymbol{u}}_1$ for the linear and Gaussian kernels. Here, we used $\gamma = d^{2^s/8},\ s \in \{1, \ldots, 4\}$ and $\hat{\gamma}(\Gamma_0)$. Note that (A-v) does not hold when $\gamma = d^{2^s/8},\ s \in \{1, 2\}$, for (e) to (h). We checked the consistency (13). Note that the consistency is equivalent to $(\|\boldsymbol{r}\|^{-1}\boldsymbol{r}^\top\hat{\boldsymbol{u}}_1 - 1)^2 = o_P(1)$. Let $\hat{\boldsymbol{u}}_{1t}$ be $\hat{\boldsymbol{u}}_1$ in the tth iteration. We calculated the squared error, $\mathrm{SE}(\hat{\boldsymbol{u}}_{1t}) = (\|\boldsymbol{r}\|^{-1}\boldsymbol{r}^\top\hat{\boldsymbol{u}}_{1t} - 1)^2$, for $t \in \{1, \ldots, 2000\}$. We repeated it 2000 times and took the average, $\mathrm{MSE}(\hat{\boldsymbol{u}}_1) = \sum_{t=1}^{2000}\mathrm{SE}(\hat{\boldsymbol{u}}_{1t})/2000$. In Fig. 4, we plotted $\mathrm{MSE}(\hat{\boldsymbol{u}}_1)$ for $d = 2^s,\ s \in \{3, \ldots, 11\}$. We observed that the Gaussian kernel with $\hat{\gamma}(\Gamma_0)$ gives preferable performances for (e) to (h). The performances of the Gaussian kernel with $\gamma = d^2$ were close to those of the linear kernel. See Proposition 5.

Next, we compare the performance of the proposed methods in complex settings. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = 0.9\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = 1.1(0.4^{|i-j|^{1/3}})$. Note that $\Delta_\Sigma = 0.2d$. For Π_is, we considered the following distributions:

(A) $\Pi_i : \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$;
(B) $(z_{i1}, \ldots, z_{id})^\top$s are i.i.d. as the d-variate t-distribution, $t_d(\boldsymbol{I}_d, \nu)$, with mean zero, covariance matrix $\boldsymbol{I}_d$ and degrees of freedom $\nu = 5$;
(C) $z_{i\ell} = (v_{i\ell} - 5)/\sqrt{10}$ $(\ell = 1, \ldots, d)$ in which $v_{i\ell}$s are i.i.d. as the chi-squared distribution with 5 degrees of freedom.

Here, $z_{i\ell}$s are given in Remark 2.

Note that (A-i) holds for (A) and (C), however does not hold for (B). We set up three cases:

(i) $n_1 = n_2 = 8 + 4s,\ s \in \{1, \ldots, 9\}$, $\boldsymbol{\mu}_2 = \boldsymbol{1}_d/5$, and $d = 1000$ for (A);
(j) $n_1 = 2s, n_2 = 5s,\ s \in \{1, \ldots, 9\}$, $\boldsymbol{\mu}_2 = \boldsymbol{1}_d$, and $d = 1000$ for (B);
(k) $n_1 = n_2 = 10$, $\boldsymbol{\mu}_2 = \boldsymbol{1}_d/3$, and $d = 2^s,\ s \in \{3, \ldots, 11\}$ for (C).

Similar to the above simulations, we calculated $\mathrm{MSE}(\hat{\boldsymbol{u}}_1)$ for the kernel functions (I) and (II) with $\gamma = d^{2^s/8}$ $(s \in \{1, \ldots, 4\})$ and $\hat{\gamma}(\Gamma_0)$ and plotted the results in Fig. 5. We observed that the Gaussian kernel with $\hat{\gamma}(\Gamma_0)$ gives adequate performances even when n_is are large or unbalanced. Also, it gave preferable performances for the non-Gaussian cases.

5.2. Data examples

In this section, we analyzed three microarray data sets given in the supplemental material of Mramor et al. [19]. See the web page (http://www.biolab.si/supp/bi-cancer/projections/index.html) for the details. The three data sets are as follows:

(D-i) Lung cancer data with 12 600 genes, consisting of Π_1: normal lung (17 samples) and Π_2: small cell lung cancer (6 samples), given by Bhattacharjee et al. [9];

(D-ii) Leukemia data with 12 533 genes, consisting of Π_1: mixed-lineage leukemia (20 samples) and Π_2: acute myeloid leukemia (28 samples), given by Armstrong et al. [8];

(D-iii) Lymphoma and leukemia data with 15 434 genes, consisting of Π_1: T-cell lymphoblastic lymphoma (9 samples) and Π_2: B-cell acute lymphoblastic leukemia (10 samples), given by Raetz et al. [21].

In Fig. 6, we displayed scatter plots of $(s_{1j(I)}, s_{2j(I)})$ and $(s_{1j(II)}, s_{2j(II)})$ with $\gamma = \hat{\gamma}(\Gamma_0)$, $j \in \{1, \ldots, n\}$, together with two vertical lines, $\sqrt{n_2/n_1}$ and $-\sqrt{n_1/n_2}$, for each data set. Also, we gave the value of $\mathrm{SE}(\hat{\boldsymbol{u}}_1)$ for the linear and Gaussian kernels in Table 1. From Fig. 6, one can effectively cluster $\boldsymbol{x}_j$s into two groups by the sign of the PC scores. Especially, for (D-ii), $s_{ij(II)}$s gave a better performance than $s_{ij(I)}$s. In fact, $\mathrm{SE}(\hat{\boldsymbol{u}}_1)$ was smaller for the Gaussian kernel compared with the linear kernel.

Table 1
The value of $\mathrm{SE}(\hat{\boldsymbol{u}}_1)$ for three microarray data sets, (D-i), (D-ii) and (D-iii), in cases of (I) the linear kernel and (II) the Gaussian kernel.

(n_1, n_2)	$\mathrm{SE}(\hat{\boldsymbol{u}}_1)$ for (I)	$\mathrm{SE}(\hat{\boldsymbol{u}}_1)$ for (II)
(D-i) (17, 6)	0.001	0.003
(D-ii) (20, 28)	0.051	0.018
(D-iii) (9, 10)	0.016	0.006

6. Conclusion

In this paper, we considered a clustering method based on the KPCA for HDLSS data. We first investigated asymptotic properties of the KPCA with the linear and Gaussian kernels for the two-class ($k = 2$) model. We theoretically showed that HDLSS data can be classified by the sign of the first PC scores. See also Appendix C of the online supplementary for the case when $k = 3$. Detailed study of the case when $k \geq 4$ is left to a future work. We gave theoretical reasons why the Gaussian kernel is effective for clustering high-dimensional data. We discussed the choice of the scale parameter, γ, to enjoy high performances of the KPCA with the Gaussian kernel. We showed that the Gaussian kernel with the γ gives preferable performances both in numerical simulations and actual data analyses. However, we have to say, the dataset is not always classified by the sign of the first several PC scores. See Fig. 6 or Fig. C.2 in Appendix C of the online supplementary. Therefore, we recommend the following steps: (i) apply the KPCA with the Gaussian kernel, (ii) map the dataset onto the first two or three eigenspaces (feature space), and (iii) apply general clustering methods such as the k-means method to the feature space.

CRediT authorship contribution statement

Yugo Nakayama: Methodology, Software, Writing - original draft. **Kazuyoshi Yata:** Conceptualization, Methodology, Investigation, Writing - original draft. **Makoto Aoshima:** Methodology, Supervision, Project administration, Funding acquisition, Writing - review & editing.

Acknowledgments

We would like to thank the two anonymous referees for their constructive comments. We are also grateful to the Editor-in-Chief, Dietrich von Rosen, for his editorial assistance. The research of the second author was partially supported by Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS), under Contract Number 18K03409. The research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 20H00576 and 19K22837.

Appendix A. General framework of the kernel PCA

In this section, we consider the KPCA in a general framework. We assume the following condition as $d \to \infty$:

(A-ix) $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \delta_i + o_P(\Delta)$ when $\boldsymbol{x}_j, \boldsymbol{x}_{j'} \in \Pi_i$ $(j \neq j')$, $i \in \{1, 2\}$,
$k(\boldsymbol{x}_j, \boldsymbol{x}_j) = \delta_{2+i} + o_P(\Delta)$ when $\boldsymbol{x}_j \in \Pi_i$, $i \in \{1, 2\}$,
and $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \delta_5 + o_P(\Delta)$ when $\boldsymbol{x}_j \in \Pi_1$, $\boldsymbol{x}_{j'} \in \Pi_2$,
where $\Delta = \delta_1 + \delta_2 - 2\delta_5$ and δ_ls are variables (which may depend on d) such that $\Delta > 0$, $\delta_3 \geq \delta_1$ and $\delta_4 \geq \delta_2$.

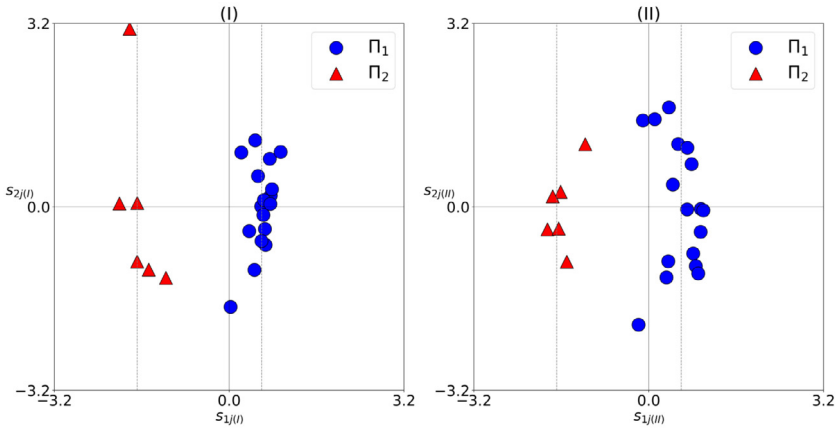
Let $\sigma_1 = \delta_3 - \delta_1$, $\sigma_2 = \delta_4 - \delta_2$ and $\Delta_\sigma = |\sigma_2 - \sigma_1|$. Note that (A-ix) is regarded as a convergence condition for the gram matrix and Δ is a distance between the two populations. Also, note that δ_is are characteristic variables for each kernel in high-dimensional settings. For example, $\Delta = \kappa_1^2 + \kappa_2^2 - 2\kappa_1\kappa_2\kappa_\mu = \Delta_\kappa \kappa_1^2$, $\delta_i = \kappa_i^2$, $i \in \{1, 2\}$, $\delta_3 = \delta_4 = 1$, $\delta_5 = \kappa_1\kappa_2\kappa_\mu$ and $\Delta_\sigma = \kappa_1^2 - \kappa_2^2 = \kappa_1^2(1 - \kappa_\Sigma^2)$ when $k(\cdot, \cdot)$ is the Gaussian kernel function (II). Also, from (B.3) and (B.4) in Appendix B, (A-ix) is met for the Gaussian kernel under (A-v) and (A-vi).

Remark 8. We note that $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, $\delta_i = \|\boldsymbol{\mu}_i\|^2$, $\delta_{2+i} = \|\boldsymbol{\mu}_i\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_i)$, $i \in \{1, 2\}$, $\delta_5 = \boldsymbol{\mu}_1^\top \boldsymbol{\mu}_2$ and $\Delta_\sigma = \Delta_\Sigma$ when $k(\cdot, \cdot)$ is the linear kernel function (I). See Nakayama et al. [20] for δ_is of the polynomial kernel function (III).

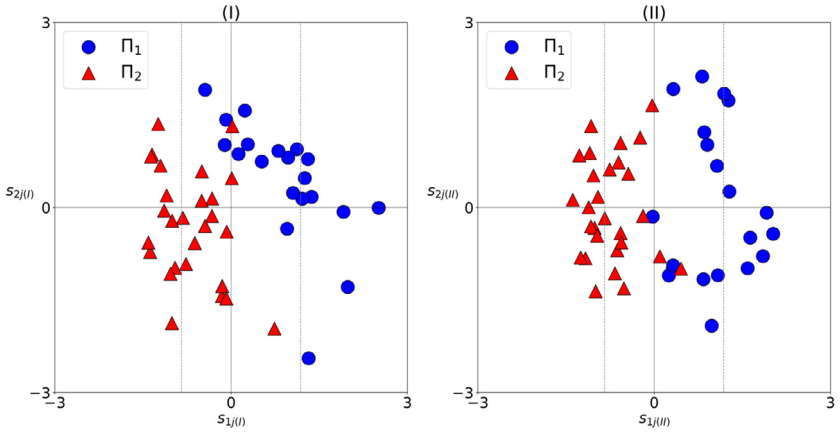
For the sake of simplicity, we assume $\sigma_1 \leq \sigma_2$. Let

$$\widetilde{\boldsymbol{K}}_0 = \Delta \boldsymbol{r}\boldsymbol{r}^\top + \boldsymbol{P}_n \begin{pmatrix} \sigma_1 \boldsymbol{I}_{n_1} & \boldsymbol{O}_{n_1, n_2} \\ \boldsymbol{O}_{n_2, n_1} & \sigma_2 \boldsymbol{I}_{n_2} \end{pmatrix} \boldsymbol{P}_n.$$

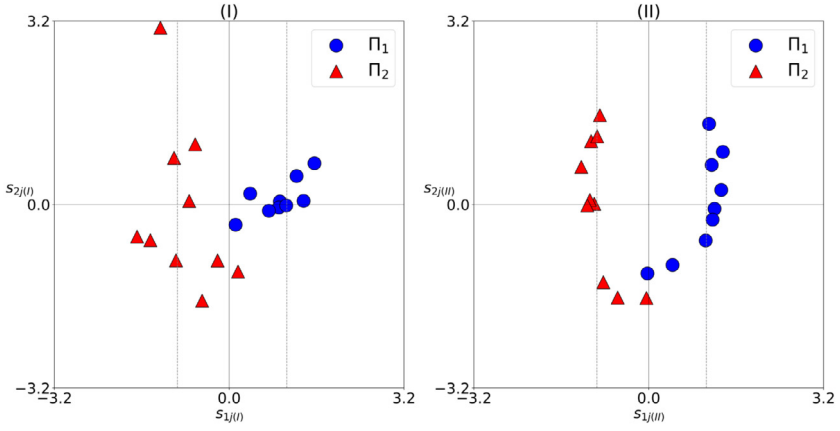
Then, under (A-ix), it holds that $\|\boldsymbol{K}_0 - \widetilde{\boldsymbol{K}}_0\|_F = o_P(\Delta)$ as $d \to \infty$. Thus, we have the following result.



(D-i) Bhattacharjee et al. [9]: $(n_1, n_2) = (17, 6)$.



(D-ii) Armstrong et al. [8]: $(n_1, n_2) = (20, 28)$.



(D-iii) Raetz et al. [21]: $(n_1, n_2) = (9, 10)$.

Fig. 6. Scatter plots of the PC scores for (D-i) to (D-iii). The blue circles and red triangles denote the data points belonging to Π_1 and Π_2, respectively. The left (right) panels illustrate the linear (Gaussian) kernel.

Proposition 6. *Assume (A-ix). Assume also*

$$\limsup_{d\to\infty} \frac{\Delta_\sigma}{n_1 \Delta} < 1 \quad \textit{when } n_2 \geq 2.$$

Then, it holds as $d \to \infty$

$$s_{1j} = \begin{cases} \sqrt{n_2/n_1} + o_P(1), & j \in \{1, \ldots, n_1\}, \\ -\sqrt{n_1/n_2} + o_P(1), & j \in \{n_1+1, \ldots, n\}. \end{cases}$$

In addition, it holds as $d \to \infty$

$$\sum_{j'=n_1+1}^{n} \frac{s_{ij'}^2}{n} = 1 + o_P(1), \quad s_{ij} = o_P(1),\ j \in \{1, \ldots, n_1\},\ i \in \{2, \ldots, n_2\}$$

under $n_2 \geq 2$ *and* $\liminf_{d\to\infty} \Delta_\sigma/\Delta > 0$.

We checked the performance of the PC scores by the polynomial and Laplace kernels in the same settings (a) and (d) as in Section 2.2. We set $(\zeta, r) = (d, 2)$ in (III) and $\xi = d$ in (IV). Let $s_{ij(III)}$ and $s_{ij(IV)}$ denote s_{ij} given by using the kernel functions (III) and (IV), respectively. In Fig. A.1, we displayed scatter plots of $(s_{1j(III)}, s_{2j(III)})$ and $(s_{1j(IV)}, s_{2j(IV)})$, $j \in \{1, \ldots, n\}$, together with two theoretical lines, $\sqrt{n_2/n_1}$ and $-\sqrt{n_1/n_2}$. We observed that both the kernels give good performances for (a). For (d), $s_{1j(IV)}$s became close to the theoretical lines as d increases. However, $s_{1j(III)}$s did not show the consistency property for (d). This is probably because the Laplace kernel can draw information about heteroscedasticity via the difference of $\boldsymbol{\Sigma}_i$s.

Appendix B. Proofs

Let $\boldsymbol{L}_1 = \begin{pmatrix} \mathrm{tr}(\boldsymbol{\Sigma}_1)\boldsymbol{I}_{n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & \mathrm{tr}(\boldsymbol{\Sigma}_2)\boldsymbol{I}_{n_2} \end{pmatrix}$ and $\boldsymbol{L}_2 = \begin{pmatrix} (1-\kappa_1^2)\boldsymbol{I}_{n_1} & \boldsymbol{O}_{n_1,n_2} \\ \boldsymbol{O}_{n_2,n_1} & (1-\kappa_1^2\kappa_\Sigma^2)\boldsymbol{I}_{n_2} \end{pmatrix}$.

Let $\boldsymbol{J}_{n_1,n_2}$ be the $n_1 \times n_2$ matrix with all elements 1.

Proof of Lemma 1. Let $\boldsymbol{\mu}_0 = \eta_1\boldsymbol{\mu}_1 + \eta_2\boldsymbol{\mu}_2$. We can write that $\boldsymbol{x}_j - \boldsymbol{\mu}_0 = (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + (-1)^{i+1}(1-\eta_i)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for $i \in \{1, 2\}$, $j \in \{1, \ldots, n\}$. Under (A-i) and (A-ii), it holds that as $d \to \infty$ for $i \in \{1, 2\}$

$$\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 | \boldsymbol{x} \in \Pi_i)/\Delta_\mu^2 = o(1).$$

Then, similar to the proof of Lemma 1 in Yata and Aoshima [28], under (A-i) and (A-ii), we have that as $d \to \infty$

$$\begin{aligned} &\{\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)\}/\Delta_\mu = o_P(1), \quad (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_{i'})/\Delta_\mu = o_P(1) \\ &\text{and} \quad (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\Delta_\mu = O_P[\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}^{1/2}/\Delta_\mu] = o_P(1) \end{aligned} \tag{B.1}$$

when $\boldsymbol{x}_j \in \Pi_i$ and $\boldsymbol{x}_{j'} \in \Pi_{i'}$ $(j \neq j')$, from the fact that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leq \Delta_\mu \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$. Thus, it holds that

$$\left\|(\boldsymbol{X} - \boldsymbol{\mu}_0\boldsymbol{1}_n^\top)^\top(\boldsymbol{X} - \boldsymbol{\mu}_0\boldsymbol{1}_n^\top) - \Delta_\mu\boldsymbol{r}\boldsymbol{r}^\top - \boldsymbol{L}_1\right\|_F = o_P(\Delta_\mu)$$

under (A-i) and (A-ii). By noting that $\boldsymbol{P}_n(\boldsymbol{X} - \boldsymbol{\mu}_0\boldsymbol{1}_n^\top)^\top(\boldsymbol{X} - \boldsymbol{\mu}_0\boldsymbol{1}_n^\top)\boldsymbol{P}_n = \boldsymbol{K}_{0(I)}$ and $\boldsymbol{r}^\top\boldsymbol{P}_n = \boldsymbol{r}^\top$ from $\boldsymbol{r}^\top\boldsymbol{1}_n = 0$, we conclude the result.

Proof of Theorem 1, Propositions 1 and 2. Assume (A-i) and (A-ii). Note that $\hat{\boldsymbol{u}}_i^\top\boldsymbol{1}_n = 0$ when $\hat{\lambda}_i > 0$ since $\boldsymbol{1}_n^\top\boldsymbol{K}_{0(I)}\boldsymbol{1}_n = 0$. From Lemma 1, we have that as $d \to \infty$

$$\frac{\hat{\boldsymbol{u}}_i^\top\boldsymbol{K}_{0(I)}\hat{\boldsymbol{u}}_i}{\Delta_\mu} = (\hat{\boldsymbol{u}}_i^\top\boldsymbol{r})^2 + \frac{\hat{\boldsymbol{u}}_i^\top\boldsymbol{P}_n\boldsymbol{L}_1\boldsymbol{P}_n\hat{\boldsymbol{u}}_i}{\Delta_\mu} + o_P(1) = (\hat{\boldsymbol{u}}_i^\top\boldsymbol{r})^2 + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)}{\Delta_\mu} + \frac{\Delta_\Sigma\hat{\boldsymbol{u}}_i^\top\boldsymbol{P}_n\boldsymbol{D}_n\boldsymbol{P}_n\hat{\boldsymbol{u}}_i^\top}{\Delta_\mu} + o_P(1) \tag{B.2}$$

when $\hat{\lambda}_i > 0$, where $\boldsymbol{D}_n = \mathrm{diag}(0, \ldots, 0, 1, \ldots, 1)$ whose last n_2 diagonal elements are 1. When $n_2 \geq 2$, from (6), under (A-iii), it holds that $\hat{\boldsymbol{u}}_1^\top\boldsymbol{r}/\|\boldsymbol{r}\| = 1 + o_P(1)$ since $(\boldsymbol{1}_{n_1}^\top, -\boldsymbol{1}_{n_2}^\top)\hat{\boldsymbol{u}}_1 \geq 0$. When $n_2 = 1$ and $\hat{\boldsymbol{u}}_1^\top\boldsymbol{1}_n = 0$, we note that

$$\underset{\hat{\boldsymbol{u}}_1}{\mathrm{argmax}}(\hat{\boldsymbol{u}}_1^\top\boldsymbol{P}_n\boldsymbol{D}_n\boldsymbol{P}_n\hat{\boldsymbol{u}}_1) = \boldsymbol{r}/\|\boldsymbol{r}\|.$$

Thus it concludes the result of Theorem 1.

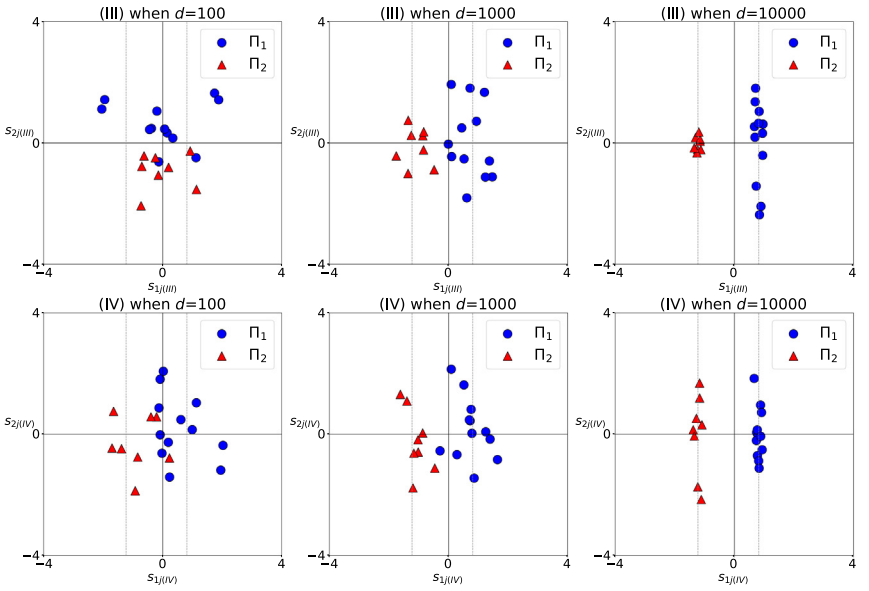
For Proposition 1, from (5), under (A-iii) and (A-iv), it hold that

$$\hat{\boldsymbol{u}}_i = (o_P(1), \ldots, o_P(1), \hat{u}_{in_1+1}, \ldots, \hat{u}_{in})^\top$$

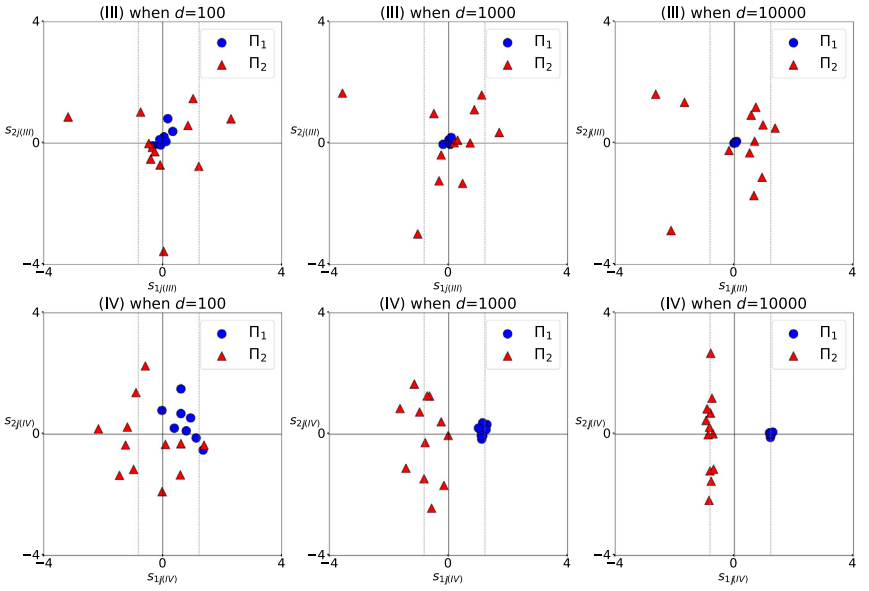
for $i \in \{2, \ldots, n_2\}$. It concludes the result of Proposition 1. Similarly, from (5) and (B.2), we can conclude the result of Proposition 2.

Proof of Lemma 2. Assume (A-i) and (A-v). Let $\omega = \{\max_{i=1,2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \max_{i=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}^{1/2}/\gamma$. Note that $\omega = o(1)$ as $d \to \infty$ under (A-vi) from the fact that $\Delta_\kappa = O(1)$. Thus from (B.1), under (A-vi), it holds that

$$\begin{aligned} &\exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma) = \exp(-2\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma)\{1 + O_P(\omega)\}, \boldsymbol{x}_j, \boldsymbol{x}_{j'} \in \Pi_i\ (j \neq j'); \\ &\exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma) = \exp\{-(\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \Delta_\mu)/\gamma\}\{1 + O_P(\omega)\}, \boldsymbol{x}_j \in \Pi_1, \boldsymbol{x}_{j'} \in \Pi_2. \end{aligned} \tag{B.3}$$



(a) When $\boldsymbol{\mu}_1 = 3^{-1}\mathbf{1}_d$, $c_2 = 1$ and $(n_1, n_2) = (12, 8)$.



(d) When $\boldsymbol{\mu}_1 = 3^{-1}\mathbf{1}_d$, $c_2 = 2$ and $(n_1, n_2) = (8, 12)$.

Fig. A.1. Toy example to illustrate the behaviors of the PC scores by the polynomial and Laplace kernels for (a) and (d). The blue circles and red triangles denote the data points belonging to Π_1 and Π_2, respectively. The upper (lower) three panels illustrate the polynomial (Laplace) kernel for (a) and (d).

Thus, under (A-vi), it holds that

$$\left\| \boldsymbol{K}/\kappa_1^2 - \begin{pmatrix} \boldsymbol{J}_{n_1,n_1} & \kappa_\mu \kappa_\Sigma \boldsymbol{J}_{n_1,n_2} \\ \kappa_\mu \kappa_\Sigma \boldsymbol{J}_{n_2,n_1} & \kappa_\Sigma^2 \boldsymbol{J}_{n_2,n_2} \end{pmatrix} - \boldsymbol{L}_2/\kappa_1^2 \right\|_F = O_P(\omega). \tag{B.4}$$

Note that

$$\boldsymbol{P}_n \begin{pmatrix} \boldsymbol{J}_{n_1,n_1} & \kappa_\mu \kappa_\Sigma \boldsymbol{J}_{n_1,n_2} \\ \kappa_\mu \kappa_\Sigma \boldsymbol{J}_{n_2,n_1} & \kappa_\Sigma^2 \boldsymbol{J}_{n_2,n_2} \end{pmatrix} \boldsymbol{P}_n = \Delta_\kappa \boldsymbol{r}\boldsymbol{r}^\top. \tag{B.5}$$

Then, from (B.4) and (B.5), we can conclude the result of Lemma 2.

Proof of Propositions 3 and 5. Note that (A-vi) holds under (11) when $\Delta_\mu/\gamma \in (0,\infty)$ or $\Delta_\Sigma/\gamma \in (0,\infty)$ as $d \to \infty$, from the fact that $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \le \Delta_\mu \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}$. Thus, we consider the case when $(\Delta_\mu+\Delta_\Sigma)/\gamma = o(1)$. It holds that $\Delta_\kappa = (1-\kappa_\Sigma)^2 + 2(1-\kappa_\mu)\kappa_\Sigma = (\Delta_\Sigma/\gamma)^2\{1+o(1)\} + 2(\Delta_\mu/\gamma)\{1+o(1)\}$. Thus (A-vi) holds under (11) when $(\Delta_\mu+\Delta_\Sigma)/\gamma = o(1)$. It concludes the result of Proposition 3.

Next, we consider Proposition 5. Note that $\Delta_\kappa = 2(\Delta_\mu/\gamma)\{1+o(1)\}$ and $1-\kappa_i = (\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma)\{1+o(1)\}$, $i \in \{1,2\}$, under the conditions in Proposition 5. Thus we can conclude the result of Proposition 5.

Proof of Theorem 2 and Proposition 4. Assume (A-i), (A-v) and (A-vi). Similar to (B.2), from Lemma 2, we have that as $d \to \infty$

$$\frac{\hat{\boldsymbol{u}}_i^\top \boldsymbol{K}_{0(II)} \hat{\boldsymbol{u}}_i}{\Delta_\kappa \kappa_1^2} = (\hat{\boldsymbol{u}}_i^\top \boldsymbol{r})^2 + \frac{\hat{\boldsymbol{u}}_i^\top \boldsymbol{P}_n \boldsymbol{L}_2 \boldsymbol{P}_n \hat{\boldsymbol{u}}_i}{\Delta_\kappa \kappa_1^2} + o_P(1) = (\hat{\boldsymbol{u}}_i^\top \boldsymbol{r})^2 + \frac{1-\kappa_1^2}{\Delta_\kappa \kappa_1^2} + (1-\kappa_\Sigma^2)\frac{\hat{\boldsymbol{u}}_i^\top \boldsymbol{P}_n \boldsymbol{D}_n \boldsymbol{P}_n \hat{\boldsymbol{u}}_i^\top}{\Delta_\kappa} + o_P(1)$$

when $\hat{\lambda}_i > 0$, where $\boldsymbol{D}_n$ is given in the proofs of Theorem 1, Propositions 1 and 2. Then, similar to the proofs of Theorem 1, Propositions 1 and 2, we can conclude the results in Theorem 2 and Proposition 4.

Proof of Proposition 6. Similar to the proofs of Theorem 1, Propositions 1 and 2, we can conclude the results in Proposition 6.

Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104779. We give asymptotic properties of the PC score with the Gaussian kernel function (II) for three classes ($k = 3$) and an R-code to calculate $s_{ij(II)}$s with $\gamma = \hat{\gamma}(\Gamma)$ in the online supplementary material.

References

[1] J. Ahn, M.H. Lee, Y.J. Yoon, Clustering high dimension, low sample size data using the maximal data piling distance, Statist. Sinica 22 (2012) 443–464.
[2] J. Ahn, J.S. Marron, K.M. Muller, Y.-Y. Chi, The high-dimension, low-sample-size geometric representation holds under mild conditions, Biometrika 94 (2007) 760–766.
[3] M. Aoshima, D. Shen, H. Shen, K. Yata, Y.-H. Zhou, J.S. Marron, A survey of high dimension low sample size asymptotics, Aust. N. Z. J. Stat. 60 (2018) 4–19.
[4] M. Aoshima, K. Yata, Two-stage procedures for high-dimensional data, Sequential Anal. (Ed. Spec. Invit. Pap.) 30 (2011) 356–399.
[5] M. Aoshima, K. Yata, A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, Ann. Inst. Stat. Math. 66 (2014) 983–1010.
[6] M. Aoshima, K. Yata, Two-sample tests for high-dimension, strongly spiked eigenvalue models, Statist. Sinica 28 (2018) 43–62.
[7] M. Aoshima, K. Yata, Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models, Ann. Inst. Stat. Math. 71 (2019) 473–503.
[8] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, S.J. Korsmeyer, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nat. Genet. 30 (2002) 41–47.
[9] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. 98 (2001) 13790–13795.
[10] P. Borysov, J. Hannig, J.S. Marron, Asymptotics of hierarchical clustering for growing dimension, J. Multivariate Anal. 124 (2014) 465–479.
[11] I.L. Dryden, Statistical analysis on high-dimensional spheres and shape spaces, Ann. Statist. 33 (2005) 1643–1665.
[12] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 427–444.
[13] K. Hellton, M. Thoresen, When and why are principal component scores a good tool for visualizing high-dimensional data?, Scand. J. Stat. 44 (2017) 581–597.
[14] H. Huang, Y. Liu, M. Yuan, J.S. Marron, Statistical significance of clustering using soft thresholding, J. Comput. Graph. Stat. 24 (2015) 975–993.
[15] S. Jung, J.S. Marron, PCA consistency in high dimension, low sample size context, Ann. Statist. 37 (2009) 4104–4130.
[16] P.K. Kimes, Y. Liu, H.D. Neil, J.S. Marron, Statistical significance for hierarchical clustering, Biometrics 73 (2017) 811–821.
[17] Z. Liu, D. Chen, H. Bensmail, Gene expression data classification with kernel principal component analysis, J. Biomed. Biotechnol. (2005) 155–159.
[18] Y. Liu, D.N. Hayes, A. Nobel, J.S. Marron, Statistical significance of clustering for high-dimension, low-sample size data, J. Amer. Statist. Assoc. 103 (2008) 1281–1293.
[19] M. Mramor, G. Leban, J. Demšar, B. Zupan, Visualization-based cancer microarray data classification analysis, Bioinformatics 23 (2007) 2147–2154.
[20] Y. Nakayama, K. Yata, M. Aoshima, Bias-corrected support vector machine with gaussian kernel in high-dimension, low-sample-size settings, Ann. Inst. Stat. Math. 72 (2020) 1257–1286.
[21] E.A. Raetz, S.L. Perkins, D. Bhojwani, K. Smock, M. Philip, W.L. Carroll, D.-J. Min, Gene expression profiling reveals intrinsic differences between T-cell acute lymphoblastic leukemia and T-cell lymphoblastic lymphoma, Pediatr. Blood Cancer 47 (2006) 130–140.
[22] F. Reverter, E. Vegas, P. Sánchez, Mining gene expression profiles: an integrated implementation of kernel principal component analysis and singular value decomposition, Genom. Proteom. Bioinf. 8 (2010) 200–210.
[23] S. Sarkar, A.K. Ghosh, On perfect clustering of high dimension, low sample size data, IEEE Trans. Pattern Anal. 42 (2020) 2257–2272.
[24] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[25] B. Schölkopf, A. Smola, K.-R. Müller, Kernel principal component analysis, in: Advances in Kernel Methods - Support Vector Learning, MIT Press, 1999, pp. 327–352.
[26] D. Shen, H. Shen, J.S. Marron, A general framework for consistency of principal component analysis, J. Mach. Learn. Res. 17 (2016) 1–34.
[27] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, J. Multivariate Anal. 105 (2012) 193–215.
[28] K. Yata, M. Aoshima, Geometric consistency of principal component scores for high-dimensional mixture models and its application, Scand. J. Stat. 47 (2020) 899–921.