이학석사학위논문

Double Data Piling for Heterogeneous Covariance Models

이질적 공분산 모형에서의 이중 데이터 파일링 현상

2023 년 8 월

서울대학교 대학원

통계학과

김 태 현

Double Data Piling for Heterogeneous Covariance Models

이질적 공분산 모형에서의 이중 데이터 파일링 현상

지도교수 정 성 규

이 논문을 이학석사 학위논문으로 제출함

2023 년 4 월

서울대학교 대학원

통계학과

김 태 현

김태현의 이학석사 학위논문을 인준함

2023 년 6 월

위 원 장 임 요 한 (인)

부위원장 정 성 규 (인)

위 원 이 권 상 (인)

Abstract

In this work, we characterize two data piling phenomenon for a high-dimensional binary classification problem with heterogeneous covariance models. The data piling refers to the phenomenon where projections of the training data onto a direction vector have exactly two distinct values, one for each class. This first data piling phenomenon occurs for any data when the dimension p is larger than the sample size n. We show that the second data piling phenomenon, which refers to a data piling of independent test data, can occur in an asymptotic context where p grows while n is fixed. We further show that a second maximal data piling direction, which gives an asymptotic maximal distance between the two piles of independent test data, can be obtained by projecting the first maximal data piling direction onto the nullspace of the common leading eigenspace. Based on the second data piling phenomenon, we propose novel linear classification rules which ensure perfect classification of high-dimension low-sample-size data under generalized heterogeneous spiked covariance models.

Keywords: High dimension low sample size, Classification, Maximal data piling, Spiked covariance model, High-dimensional asymptotics

Student Number: 2021-29052

Contents

Chapter 1

Introduction

High-Dimension Low-Sample-Size (HDLSS) data have often been found in many of scientific fields, such as microarray gene expression analysis, chemometrics, and image processing. Such HDLSS data are oftentimes best classified by linear classifiers since the dimension of data p is much larger than the sample size n. For binary classification with $p > n$, Ahn and Marron [2010] observed the data piling phenomenon, that is, projections of the training data onto a direction vector w are identical for each class. Among such directions exhibiting data piling, the *maximal data piling* direction uniquely gives the largest distance between the two piles of training data. The maximal data piling direction is defined as

$$w_{\text{MDP}} = \underset{w:\|w\|=1}{\operatorname{argmax}} \left(w^\top \mathbf{S}_B w \right) \text{ subject to } w^\top \mathbf{S}_W w = 0,$$

where $\mathbf{S}_W$ is the $p \times p$ within-class scatter matrix and $\mathbf{S}_B$ is the $p \times p$ between-class scatter matrix of training dataset $\mathcal{X}$. Ahn and Marron [2010] observed that a classification rule using w_{MDP} as the normal vector to a discrimina-

tive hyperplane achieves better classification performance than classical linear classifiers when there are significantly correlated variables.

However, the maximal data piling direction has not been considered as an appropriate classifier since it depends too much on training data, resulting in poor generalization performances [Marron et al., 2007, Lee et al., 2013]. In general, while the training data are piled on w_{MDP}, independent test data are not piled on w_{MDP}. Recently, Chang et al. [2021] revealed the existence of the *second data piling* direction, which gives a data piling of independent test data, under the HDLSS asymptotic regime of Hall et al. [2005] where the dimension of data p tends to grow while the sample size n is fixed. In addition, they showed that a negatively ridged linear discriminant vector, projected onto a low-dimensional subspace, can be a *second maximal data piling* direction, which yields a maximal asymptotic distance between two piles of independent test data.

A second data piling direction is defined asymptotically as $p \to \infty$, unlike the first data piling of training dataset $\mathcal{X}$ for any fixed $p > n$. For a sequence of directions $\{w\} = (w^{(1)}, \ldots, w^{(p-1)}, w^{(p)}, w^{(p+1)}, \ldots)$, in which $w^{(q)} \in \mathbb{R}^q$ for $q \in \mathbb{N}$, we write $w \in \mathbb{R}^p$ for the pth element of $\{w\}$. Let Y, Y' be independent random vectors from the same population of $\mathcal{X}$, and write $\pi(Y) = k$ if Y belongs to class k. Chang et al. [2021] defined the collection of all sequences of second data piling directions as

$$\mathcal{A} = \left\{ \{w\} \in \mathfrak{W}_X : \forall Y, Y' \text{ with } \pi(Y) = \pi(Y'),\ p^{-1/2} w^\top (Y - Y') \xrightarrow{P} 0 \text{ as } p \to \infty \right\}$$

where $\mathfrak{W}_X = \{\{w\} : w \in \mathcal{S}_X, \|w\| = 1 \text{ for all } p\}$, and $\mathcal{S}_X = \mathrm{span}(\mathbf{S}_W) \cup \mathrm{span}(\mathbf{S}_B)$ is the sample space. Furthermore, among the sequences of second data piling directions in $\mathcal{A}$, if $\{v\} \in \mathcal{A}$ satisfies

$$\{v\} \in \underset{\{w\} \in \mathcal{A}}{\mathrm{argmax}}\, D(w),$$

where $D(w)$ is the probability limit of $p^{-1/2}|w^\top(Y_1 - Y_2)|$ for $\pi(Y_k) = k$ ($k = 1, 2$), then we call v a *second maximal data piling* direction. Note that a second maximal data piling direction does not uniquely exist as opposed to w_{MDP}: For $\{v_1\} \in \mathcal{A}$ satisfying $D(w) \leq D(v_1)$ for any $\{w\} \in \mathcal{A}$, if $\|v_1 - v_2\| \xrightarrow{P} 0$ as $p \to \infty$ for some $\{v_2\} \in \mathcal{A}$, then $\{v_2\}$ also satisfies $D(w) \leq D(v_2)$ for any $\{w\} \in \mathcal{A}$.

Chang et al. [2021] showed that the second maximal data piling direction exists and by using such a direction, asymptotic perfect classification of independent test data is possible. They assumed that the population mean difference is as large as $\|\boldsymbol{\mu}_{(1)} - \boldsymbol{\mu}_{(2)}\| = O(p^{1/2})$ and each of two populations has a *homogeneous* spiked covariance matrix. The spiked covariance model, first introduced by Johnstone [2001], refers to high-dimensional population covariance matrix structures in which a few eigenvalues of the matrix are much larger than the other nearly constant eigenvalues [Ahn et al., 2007, Jung and Marron, 2009, Shen et al., 2016].

With such assumptions, Chang et al. [2021] showed that if $\boldsymbol{\Sigma}$ has m strong spikes, that is, m eigenvalues increase at the order of p as $p \to \infty$ while the other eigenvalues are nearly constant, averaging to $\tau^2 > 0$, then projections of independent test data tend to be respectively distributed along two parallel affine subspaces in a low-dimensional subspace $\mathcal{S} = \mathrm{span}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m, w_{\mathrm{MDP}}) \subset \mathcal{S}_X$, where $\hat{\mathbf{u}}_i$ is the ith eigenvector of $\mathbf{S}_W$. See Figure 1.1 for an illustration. Furthermore, they showed that v_α, which is obtained by projecting a ridged linear discrimination vector onto $\mathcal{S}$, is asymptotically orthogonal to these affine subspaces when the negative ridge parameter $\alpha := -\tau^2$ is used. Figure 1.1 displays that the projections of independent test data onto $v_{-\tau^2}$ are asymptotically piled on two distinct points, one for each class.

While Chang et al. [2021] provided compelling insights on double data pil-

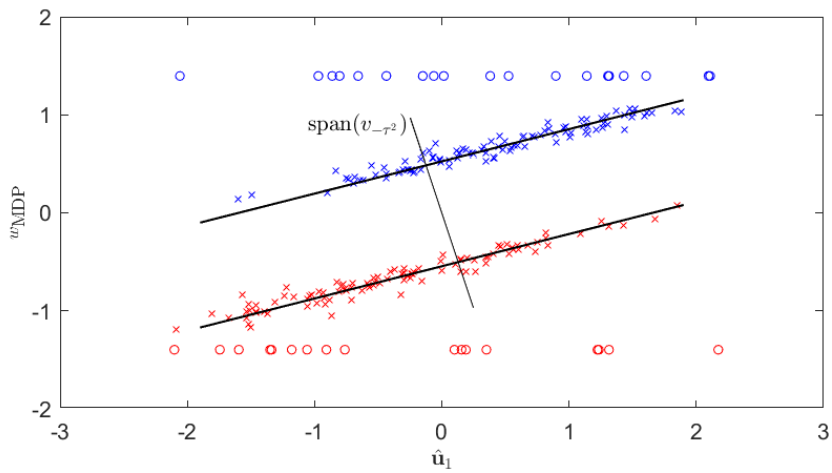

Figure 1.1 Double data piling phenomenon for homogeneous covariance model with one strong spike ($m = 1$). The projections of training dataset are piled on two distinct points on w_{MDP}. The projections of independent test dataset are distributed along parallel lines in $\mathcal{S} = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$, which appear to be orthogonal to $v_{-\tau^2}$.

ing phenomenon, their discussion was limited to the homogeneous covariance models. Also, it is known that v_α, the projected ridged linear discriminant vector, may not yield second data piling for any ridge parameter $\alpha \in \mathbb{R}$ under heterogeneous covariance models. In this work, we show that, under generalized *heterogeneous* spiked covariance models, the second data piling phenomenon occurs when the dimension of data p grows while the sample size n is fixed, and a second maximal data piling direction can be also obtained purely from the training data. Moreover, we introduce novel algorithms which ensure perfect classification of independent test data for heterogeneous covariance models, by noting the fact that a second maximal data piling direction can be obtained by projecting w_{MDP} onto the nullspace of the common leading eigenspace.

The rest of this paper is organized as follows. In Chapter 2, we specifically define the generalized heterogeneous spiked covariance models. In Chapter 3, we characterize the second data piling phenomenon under the heterogeneous covariance models. In Chapter 4, we propose Second Maximal Data Piling (SMDP) algorithms to estimate a second maximal data piling direction. In Chapter 5, we numerically confirm classification performances of SMDP algorithms. In Chapter 6, we conclude the paper with a discussion. The proofs of main lemmas and theorems are contained in Appendices A and B.

Chapter 2

Heterogeneous Covariance Models

We assume that for $k = 1, 2$, $X|\pi(X) = k$ follows an absolutely continuous distribution on $\mathbb{R}^p$ with mean $\boldsymbol{\mu}_{(k)}$ and covariance matrix $\boldsymbol{\Sigma}_{(k)}$. Also, we assume $\mathbb{P}(\pi(X) = k) = \pi_k$, where $\pi_k > 0$ and $\pi_1 + \pi_2 = 1$. Write the eigen-decomposition of $\boldsymbol{\Sigma}_{(k)}$ by $\boldsymbol{\Sigma}_{(k)} = \mathbf{U}_{(k)}\boldsymbol{\Lambda}_{(k)}\mathbf{U}_{(k)}^\top$, where $\boldsymbol{\Lambda}_{(k)} = \mathrm{Diag}(\lambda_{(k),1}, \ldots, \lambda_{(k),p})$ in which the eigenvalues are arranged in descending order, and $\mathbf{U}_{(k)} = [\mathbf{u}_{(k),1}, \ldots, \mathbf{u}_{(k),p}]$ for $k = 1, 2$.

Let the $p \times n$ data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ where $\mathbf{X}_k = [X_{k1}, \ldots, X_{kn_k}]$, $n := n_1 + n_2$ and $\pi(X_{kj}) = k$ for any k, j. We assume n_1 and n_2 are fixed and denote $\eta_k = n_k/n$ for $k = 1, 2$. We assume n_1 and n_2 are fixed and denote $\eta_k = n_k/n$ for $k = 1, 2$. We write class-wise sample mean vectors $\bar{X}_k = {n_k}^{-1} \sum_{j=1}^{n_k} X_{kj}$, and total sample mean vector $\bar{X} = \eta_1\bar{X}_1 + \eta_2\bar{X}_2$. Also, we write the within-class scatter matrix $\mathbf{S}_W = (\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^\top$ where $\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1\ \bar{\mathbf{X}}_2]$ and $\bar{\mathbf{X}}_k = \bar{X}_k \mathbf{1}_{n_k}^\top$ for $k = 1, 2$. We write an eigen-decomposition of $\mathbf{S}_W$ by $\mathbf{S}_W = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{U}}^\top$, where $\hat{\boldsymbol{\Lambda}} = \mathrm{Diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ in which the eigenvalues are arranged in descending order, and $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_p]$. Since $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{n-2} \geq \hat{\lambda}_{n-1} = \ldots = \hat{\lambda}_p = 0$

with probability 1, we can write $\mathbf{S}_W = \hat{\mathbf{U}}_1 \hat{\mathbf{\Lambda}}_1 \hat{\mathbf{U}}_1^\top$ where $\hat{\mathbf{U}}_1 = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n-2}]$ and $\hat{\mathbf{\Lambda}}_1 = \mathrm{Diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{n-2})$. Also, we write $\hat{\mathbf{U}}_2 = [\hat{\mathbf{u}}_{n-1}, \ldots, \hat{\mathbf{u}}_p]$. We denote the sample space as $\mathcal{S}_X$, which is the $(n-1)$-dimensional subspace spanned by $X_{kj} - \bar{X}$ for $k = 1, 2$ and $1 \leq j \leq n_k$. Note that the sample space $\mathcal{S}_X$ can be equivalently expressed as $\mathrm{span}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n-2}, w_{\mathrm{MDP}})$ [Ahn and Marron, 2010, Chang et al., 2021]. We denote the sample mean difference vector as $\mathbf{d} = \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$. Note that the sphered data matrix of $\mathbf{X}_k$ is $\mathbf{Z}_{(k)} = \mathbf{\Lambda}_{(k)}^{-\frac{1}{2}} \mathbf{U}_{(k)}^\top (\mathbf{X}_k - \boldsymbol{\mu}_{(k)} \mathbf{1}_{n_k}^\top) = [z_{(k),1}, \ldots, z_{(k),p}]^\top \in \mathbb{R}^{p \times n_k}$ for $k = 1, 2$. Then the elements of $\mathbf{Z}_{(k)}$ are uncorrelated with each other, and have mean zero and unit variance. We make the following assumptions for generalized heterogeneous spiked covariance models.

Assumption 1 *For the population mean difference vector $\boldsymbol{\mu} = \boldsymbol{\mu}_{(1)} - \boldsymbol{\mu}_{(2)}$, there exists $\delta > 0$ such that $p^{-1/2} \|\boldsymbol{\mu}\|_2 \to \delta$ as $p \to \infty$.*

Assumption 2 *For a fixed integer $m_k \geq 1$, $\sigma^2_{(k),i}, \tau^2_{(k),i} > 0$ $(k = 1, 2)$, assume that $\lambda_{(k),i} = \sigma^2_{(k),i} p + \tau^2_{(k),i}$ for $1 \leq i \leq m_k$ and $\lambda_{(k),i} = \tau^2_{(k),i}$ for $m_k + 1 \leq i \leq p$. Also, $\left\{ \tau^2_{(k),i} : k = 1, 2,\ i = 1, 2, \ldots \right\}$ is uniformly bounded and $p^{-1} \sum_{i=1}^p \tau^2_{(k),i} \to \tau^2_k$ as $p \to \infty$ for some $\tau^2_k > 0$.*

Assumption 1 ensures that nearly all variables are meaningfully contributing to discrimination [Hall et al., 2005, Qiao et al., 2010, Jung, 2018]. Assumption 2 allows heterogeneous covariance matrices for different classes, including the homogeneous case, that is, $\mathbf{\Sigma}_{(1)} = \mathbf{\Sigma}_{(2)}$. We assume for $k = 1, 2$, $\mathbf{\Sigma}_{(k)}$ has m_k strong spikes, that is, m_k eigenvalues increase at the order of p as $p \to \infty$ while the other eigenvalues are nearly constant as τ^2_k. We call the first m_k eigenvalues and their corresponding eigenvectors leading eigenvalues and eigenvectors of the kth class for $k = 1, 2$.

Also, we regulate the dependency of the principal components by introducing the concept of ρ-mixing condition [Kolmogorov and Rozanov, 1960, Bradley, 2005]. For any σ-field $\mathcal{E}$, denote the class of square-integrable and $\mathcal{E}$-measurable random variables as $L_2(\mathcal{E})$. Suppose $\{Z_i : -\infty \leq i \leq \infty\}$ is a sequence of random variables. For $-\infty \leq J \leq L \leq \infty$, denote $\mathcal{F}_J^L$ as the σ-field of events generated by the random variables $\{Z_i : J \leq i \leq K\}$. Then, for the ρ-mixing coefficient

$$\begin{aligned}\rho(k) &:= \sup_{j\in\mathbb{Z}} \rho(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+k}^{\infty}) \\ &= \sup_{j\in\mathbb{Z}} \sup \left\{ |\mathrm{Corr}(f,g)| : f \in \mathcal{L}^2(\mathcal{F}_{-\infty}^{j}), g \in \mathcal{L}^2(\mathcal{F}_{j+k}^{\infty}) \right\},\end{aligned}$$

the sequence $\{Z_i : -\infty \leq i \leq \infty\}$ is said to be ρ-mixing if $\rho(k) \to 0$ as $k \to \infty$. We now give a following assumption on the true principal component scores $z_{kj} = \mathbf{\Lambda}_{(k)}^{-1/2}\mathbf{U}_{(k)}^{\top}(X_{kj} - \boldsymbol{\mu}_{(k)}) \in \mathbb{R}^p$ for $k = 1, 2$ and $1 \leq j \leq n_k$. This allows us to make use of the law of large numbers applied to $p \to \infty$ introduced in Hall et al. [2005] and Jung and Marron [2009].

Assumption 3 *The elements of the p-vector z_{kj} have uniformly bounded fourth moments, and for each p, z_{kj} consists of the first p elements of an infinite random sequence*

$$(z_{(k),1}, z_{(k),2}, \ldots)_j,$$

which is ρ-mixing under some permutation.

We define $\mathrm{Angle}(w_1, w_2) := \arccos\{w_1^{\top} w_2 / (\|w_1\|_2 \|w_2\|_2)\}$ for $w_1, w_2 \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$. For $w \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$ and a subspace $\mathcal{V}$ of $\mathbb{R}^p$, let $P_{\mathcal{V}} w$ be the orthogonal projection of w onto $\mathcal{V}$ and define $\mathrm{Angle}(w, \mathcal{V}) := \arccos\{w^{\top} P_{\mathcal{V}} w / (\|w\|_2 \|P_{\mathcal{V}} w\|_2)\}$. Also, for subspaces $\mathcal{H} = \mathrm{span}(h_1, \ldots, h_k)$ and $\mathcal{V}$ of $\mathbb{R}^p$, we define the projection of $\mathcal{H}$ onto $\mathcal{V}$ as $P_{\mathcal{V}}\mathcal{H} = \mathrm{span}(P_{\mathcal{V}} h_1, \ldots, P_{\mathcal{V}} h_k)$. Assumption 4 specifies limiting angles between leading eigenvectors of each class and the population mean

difference vector $\boldsymbol{\mu}$. Without loss of generality, we assume $\mathbf{u}_{(k),i}^\top \boldsymbol{\mu} \geq 0$ for all $k = 1, 2$ and $1 \leq i \leq m_k$.

Assumption 4 *For $\theta_{(k),i} \in [0, \pi/2]$, $\text{Angle}(\mathbf{u}_{(k),i}, \boldsymbol{\mu}) \to \theta_{(k),i}$ as $p \to \infty$ for $1 \leq i \leq m_k$ and $k = 1, 2$.*

We write a $p \times m_k$ orthonormal matrix of leading eigenvectors of each class as $\mathbf{U}_{(k),1} = [\mathbf{u}_{(k),1}, \ldots, \mathbf{u}_{(k),m_k}]$ for $k = 1, 2$. We call $\mathcal{U}_{(k)} = \text{span}(\mathbf{U}_{(k),1})$ the leading eigenspace of the kth class. Furthermore, let $\mathcal{U}$ be the subspace spanned by leading eigenvectors whose corresponding eigenvalues increase at the order of p, that is, $\mathcal{U} = \text{span}(\mathcal{U}_{(1)}) \cup \text{span}(\mathcal{U}_{(2)})$. We call $\mathcal{U}$ the common leading eigenspace of both classes. We assume that the dimension of $\mathcal{U}$,

$$m = \dim(\mathcal{U}),$$

is a fixed constant for all p. Note that $\max(m_1, m_2) \leq m \leq m_1 + m_2$. Write an orthogonal basis of $\mathcal{U}$ as $\mathbf{U}_1 = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$, satisfying $\mathbf{u}_i^\top \boldsymbol{\mu} \geq 0$ for all $1 \leq i \leq m$. Then there exist orthogonal matrices $\mathbf{R}_{(k)}^{(p)} \in \mathbb{R}^{m \times m_k}$ satisfying $\mathbf{U}_{(k),1} = \mathbf{U}_1 \mathbf{R}_{(k)}^{(p)}$ for $k = 1, 2$. Note that the matrix $\mathbf{R}_{(k)}^{(p)}$ catches the angles between the m_k leading eigenvectors in $\mathbf{U}_{(k),1}$ and the m basis in $\mathbf{U}_1$. We assume the following.

Assumption 5 *For $\theta_i \in [0, \pi/2]$, $\text{Angle}(\mathbf{u}_i, \boldsymbol{\mu}) \to \theta_i$ as $p \to \infty$ for $1 \leq i \leq m$ and for an orthogonal matrix $\mathbf{R}_{(k)} \in \mathbb{R}^{m \times m_k}$, $\mathbf{R}_{(k)}^{(p)} \to \mathbf{R}_{(k)}$ as $p \to \infty$ for $k = 1, 2$. Moreover, $\mathbf{R} = [\mathbf{R}_{(1)}\ \mathbf{R}_{(2)}]$ is of rank m.*

Finally, let φ denote the limiting angle between $\boldsymbol{\mu}$ and $\mathcal{U}$. Then we have $\cos^2 \varphi = \sum_{i=1}^m \cos^2 \theta_i$.

Chapter 3

Data Piling of Independent Test Data

In this chapter, we show that independent test data, projected onto a low-dimensional signal subspace $\mathcal{S}$ of the sample space $\mathcal{S}_X$, tend to be respectively distributed along two affine subspaces as $p \to \infty$. Chang et al. [2021] showed that there are two affine subspaces, each with dimension $m = m_1 = m_2$, such that they are parallel to each other if each class has common covariance matrix, that is, $\mathbf{\Sigma}_{(1)} = \mathbf{\Sigma}_{(2)}$. We show that if $\mathbf{\Sigma}_{(1)} \neq \mathbf{\Sigma}_{(2)}$, these affine subspaces may not be parallel to each other, but there exist *parallel* affine subspaces, of greater dimension, containing each of these affine subspaces.

To illustrate this phenomenon, we first consider a simple one-component covariance model for each covariance matrix, that is, $m_1 = 1$ and $m_2 = 1$ in Assumption 2. In Chapter 3.1, this phenomenon is demonstrated under the one-component covariance model with various conditions on covariance matrices. In Chapter 3.2, we characterize the signal subspace $\mathcal{S}$, which captures important variability of independent test data, for each scenario of two covari-

ance matrices. Then we provide the main theorem (Theorem 5) that generalizes propositions in Chapter 3.1 to the cases where $m_1 \geq 1$ and $m_2 \geq 1$.

Let $\mathcal{Y}_k$ be an independent test data of the kth class whose element $Y \in \mathcal{Y}_k$ satisfies $\pi(Y) = k$ for $k = 1, 2$ and is independent to training data $\mathcal{X}$. Write $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2$.

3.1 One-component Covariance Model

We investigate the data piling phenomenon of independent test data under the one-component covariance model as follows:

$$\begin{aligned} \mathbf{\Sigma}_{(1)} &= \sigma^2_{(1),1} p \mathbf{u}_{(1),1} \mathbf{u}^\top_{(1),1} + \tau_1^2 \mathbf{I}_p; \\ \mathbf{\Sigma}_{(2)} &= \sigma^2_{(2),1} p \mathbf{u}_{(2),1} \mathbf{u}^\top_{(2),1} + \tau_2^2 \mathbf{I}_p. \end{aligned} \tag{3.1}$$

We consider various settings where both covariance matrices have either equal tail eigenvalues or unequal tail eigenvalues, and have either a common leading eigenvector or uncommon leading eigenvectors. We provide an overview of our settings in the following.

	$\mathbf{u}_{(1),1} = \mathbf{u}_{(2),1}$	$\mathbf{u}_{(1),1} \neq \mathbf{u}_{(2),1}$
$\tau_1^2 = \tau_2^2$	Example 1	Example 2
$\tau_1^2 \neq \tau_2^2$	Example 3	Example 4

First, we assume that two covariance matrices have equal tail eigenvalues, that is, $\tau_1^2 = \tau_2^2$. For the sake of simplicity, denote $\tau^2 := \tau_1^2 = \tau_2^2$.

Example 1 *We first consider the case where both classes have the common leading eigenvector, that is,* $\mathbf{u}_{(1),1} = \mathbf{u}_{(2),1} = \mathbf{u}_1$. *Note that if* $\sigma^2_{(1),1} = \sigma^2_{(2),1}$, *then this model is equivalent to the homogeneous covariance model* $\mathbf{\Sigma}_{(1)} = \mathbf{\Sigma}_{(2)}$, *studied in Chang et al. [2021].*

It turns out that the angle between $\hat{\mathbf{u}}_1$ *and* $\mathbf{u}_1$ *converges to a random quantity between* 0 *and* $\pi/2$, *while* $\hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_{n-2}$ *are strongly inconsistent with* $\mathbf{u}_1$ *in*

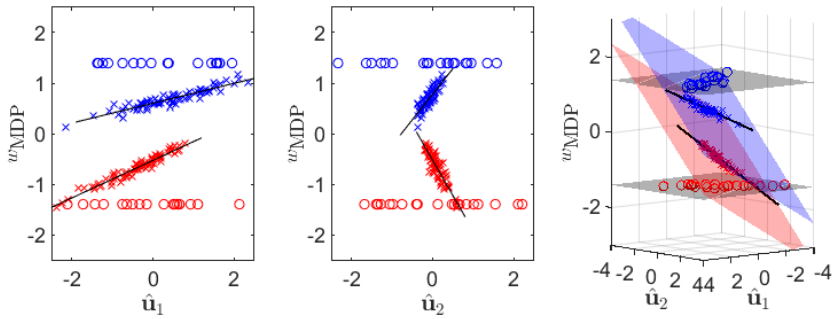

Figure 3.1 2-dimensional projections onto $\mathcal{S}_1 = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$ and $\mathcal{S}_2 = \mathrm{span}(\hat{\mathbf{u}}_2, w_{\mathrm{MDP}})$ and 3-dimensional projections onto $\mathcal{S} = \mathrm{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\mathrm{MDP}})$ with $\mathcal{D} = \{1, 2\}$ of training data $\mathcal{X}$ (class 1: blue circles, class 2: red circles) and independent test data $\mathcal{Y}$ (class 1: blue crosses, class 2: red crosses) under the model in Example 2.

the sense that $\mathrm{Angle}(\hat{\mathbf{u}}_i, \mathbf{u}_1) \xrightarrow{P} \pi/2$ *as* $p \to \infty$ *for* $2 \le i \le n-2$. *For this case, let* $\mathcal{D} = \{1\}$. *We can check that even if* $\sigma^2_{(1),1} \neq \sigma^2_{(2),1}$, *projections of independent test data* $\mathcal{Y}$ *onto* $\mathcal{S} = \mathrm{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\mathrm{MDP}}) = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$ *tend to be distributed along two parallel lines, while those of training data* $\mathcal{X}$ *are piled on two distinct points along* w_{MDP}. *This result is consistent with the findings of Chang et al. [2021] where* $\boldsymbol{\Sigma}_{(1)} = \boldsymbol{\Sigma}_{(2)}$; *see Figure 1.1. Also, the direction of these lines are asymptotically parallel to* $P_{\mathcal{S}}\mathbf{u}_1$, *which is the projection of common leading eigenvector* $\mathbf{u}_1$ *onto* $\mathcal{S}$; *see Proposition 1.*

Example 2 *Two classes do not have a common leading eigenvector, that is,* $\mathbf{u}_{(1),1} \neq \mathbf{u}_{(2),1}$, *such that the angle between* $\mathbf{u}_{(1),1}$ *and* $\mathbf{u}_{(2),1}$ *is* $\pi/4$. *Under this model, the common leading eigenspace has the dimension* $m = 2$ *(In contrast,* $m = 1$ *in the model of Example 1).*

In this case, the angle between $\hat{\mathbf{u}}_i$ *and* $\mathcal{U} = \mathrm{span}(\mathbf{u}_{(1),1}, \mathbf{u}_{(2),1})$ *converges*

to a random quantity between 0 *and* $\pi/2$ *for* $i = 1, 2$*, while the other sample eigenvectors are strongly inconsistent with* $\mathcal{U}$*. Let* $\mathcal{D} = \{1, 2\}$*. In Figure 3.1, independent test data* $\mathcal{Y}$ *projected onto* $\mathcal{S}_1 = \text{span}(\hat{\mathbf{u}}_1, w_{\text{MDP}})$ *and* $\mathcal{S}_2 = \text{span}(\hat{\mathbf{u}}_2, w_{\text{MDP}})$ *are also concentrated along lines, but in both subspaces these lines are not parallel to each other. However, within the* 3*-dimensional subspace* $\mathcal{S} = \text{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\text{MDP}}) = \text{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, w_{\text{MDP}})$*, there are two parallel* 2*-dimensional planes including these lines, one for each line. In fact,* $\mathcal{Y}_1$ *is distributed along the direction* $P_{\mathcal{S}}\mathbf{u}_{(1),1}$*, while* $\mathcal{Y}_2$ *is distributed along the direction* $P_{\mathcal{S}}\mathbf{u}_{(2),1}$*. Thus, these lines are asymptotically contained in* 2*-dimensional affine subspaces, that are parallel to* $P_{\mathcal{S}}\mathcal{U} = \text{span}(P_{\mathcal{S}}\mathbf{u}_{(1),1}, P_{\mathcal{S}}\mathbf{u}_{(2),1})$*.*

We formally state the above results. Write the scaled training data piling distance as

$$\kappa_{\text{MDP}} = p^{-1/2}\|w_{\text{MDP}}^{\top}(\bar{X}_1 - \bar{X}_2)\|. \tag{3.2}$$

For $Y \in \mathcal{Y}$ and a subspace $\mathcal{S}$ of $\mathbb{R}^p$, let $Y_{\mathcal{S}} = p^{-1/2}P_{\mathcal{S}}Y$, which is a scaled projection of Y onto $\mathcal{S}$. Similarly, write $\bar{X}_{\mathcal{S}} = p^{-1/2}P_{\mathcal{S}}\bar{X}$. Recall that $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are orthogonal basis of the common leading eigenspace $\mathcal{U} = \text{span}(\mathbf{u}_1, \ldots, \mathbf{u}_m)$. Let projections of $\mathbf{u}_i$ onto $\mathcal{S}$ as $\mathbf{u}_{i,\mathcal{S}} = P_{\mathcal{S}}\mathbf{u}_i$ and write $\mathbf{U}_{1,\mathcal{S}} = [\mathbf{u}_{1,\mathcal{S}}, \ldots, \mathbf{u}_{m,\mathcal{S}}]$. The following proposition states that for $m = 1, 2$, projections of $\mathcal{Y}$ onto the $(m+1)$-dimensional subspace $\mathcal{S} = \text{span}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m, w_{\text{MDP}})$ are distributed along two m-dimensional affine subspaces, which become parallel to each other, and also to $P_{\mathcal{S}}\mathcal{U} = \text{span}(\mathbf{u}_{1,\mathcal{S}}, \ldots, \mathbf{u}_{m,\mathcal{S}})$, as p increases.

Proposition 1 *Suppose Assumptions 1—5 hold and assume* $\tau_1^2 = \tau_2^2$ *and* $m_1 = m_2 = 1$*. Also,*

(i) if $m = 1$*, let* $\mathcal{S} = \text{span}(\hat{\mathbf{u}}_1, w_{\text{MDP}})$ *and* $L_k = \left\{\mathbf{u}_{1,\mathcal{S}}t + \nu_k w_{\text{MDP}} + \bar{X}_{\mathcal{S}} : t \in \mathbb{R}\right\}$*,*

(ii) if $m = 2$*, let* $\mathcal{S} = \text{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, w_{\text{MDP}})$ *and* $L_k = \left\{\mathbf{U}_{1,\mathcal{S}}\mathbf{t} + \nu_k w_{\text{MDP}} + \bar{X}_{\mathcal{S}} : \mathbf{t} \in \mathbb{R}^2\right\}$

for $k = 1, 2$ *where* $\nu_1 = \kappa_{\mathrm{MDP}}^{-1}\left(\eta_2(1-\cos^2\varphi)\delta^2\right)$ *and* $\nu_2 = \kappa_{\mathrm{MDP}}^{-1}\left(-\eta_1(1-\cos^2\varphi)\delta^2\right)$. *Then for any independent observation* $Y \in \mathcal{Y}$ *and for any* $\epsilon > 0$,

$$\lim_{p\to\infty} \mathbb{P}\left(\inf_{a\in L_k} \|Y_{\mathcal{S}} - a\| > \epsilon | \pi(Y) = k\right) = 0$$

for $k = 1, 2$.

Note that if $m = 1$, then $\mathcal{Y}_k$ is concentrated along the line L_k in $\mathcal{S} = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$, for $k = 1, 2$. If $m = 2$, then $\mathcal{Y}_k$ is concentrated along a line L_k', which is parallel to $P_{\mathcal{S}}\mathbf{u}_{(k),1}$ in $\mathcal{S} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, w_{\mathrm{MDP}})$, for $k = 1, 2$. Then each of the 2-dimensional subspaces L_1 and L_2 contains L_1' and L_2' respectively, and these subspaces are parallel to each other.

We now assume that two covariance matrices have unequal tail eigenvalues, that is, $\tau_1^2 \neq \tau_2^2$. Without loss of generality, we assume $\tau_1^2 > \tau_2^2$. In this case, asymptotic properties of sample eigenvectors of $\mathbf{S}_W$ are quite different from the case of $\tau_1^2 = \tau_2^2$. See Remark 1.

Remark 1 *Let* $Y = \tau_1 Y_1 U + \tau_2 Y_2(1-U) = (y_1, \ldots, y_p)^\top \in \mathbb{R}^p$, *where* Y_1, Y_2 *are two independent* $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ *random vectors,* $U = 0$ *with probability* π_2, $U = 1$ *with probability* π_1 *and* U *is independent of* Y_1, Y_2. *Note that the population covariance matrix* $\Sigma_{(0)} := Cov(Y) = (\pi_1\tau_1^2 + \pi_2\tau_2^2)\mathbf{I}_p$. *Then, the* ρ*-mixing condition for* $Z = (z_1, \ldots, z_p)^\top = \Sigma_{(0)}^{-1/2} Y$ *may or may not hold depending on whether* $\tau_1^2 = \tau_2^2$ *or not. Specifically,* Z *satisfies*

$$Cov(z_i^2, z_j^2) = \frac{1}{(\pi_1\tau_1^2 + \pi_2\tau_2^2)^2} Cov(y_i^2, y_j^2) = \frac{\pi_1\pi_2(\tau_1^2 - \tau_2^2)^2}{(\pi_1\tau_1^2 + \pi_2\tau_2^2)^2}.$$

Then in case of $\tau_1^2 = \tau_2^2$, *the sequence* $\{z_1, z_2, \ldots\}$ *is* ρ*-mixing since for all* $i \neq j$, $Cov(z_i^2, z_j^2) = 0$. *However, in case of* $\tau_1^2 \neq \tau_2^2$, *the* ρ*-mixing condition does not hold for any permuted sequence of* $\{z_1, z_2, \ldots\}$ *since* $Cov(z_i^2, z_j^2) > 0$ *for all* $i \neq j$.

This fact is relevant to different asymptotic behaviors of eigenvectors of $\mathbf{S}_W$ *depending on whether* $\tau_1^2 = \tau_2^2$ *or not. Assume that for* $k = 1, 2$, $X_{k1}, \ldots, X_{kn_k}$ *are independent* $\mathcal{N}_p(\mathbf{0}_p, \tau_k^2 \mathbf{I}_p)$ *random vectors. For the case where* $\tau_1^2 = \tau_2^2 =: \tau^2$, *Hall et al. [2005] showed that data from both classes are asymptotically located at the vertices of an n-simplex of edge length* $\sqrt{2}\tau\sqrt{p}$ *and data points tend to be orthogonal to one another, when p is extremely large. Hence, the sample eigenvectors* $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n-2}$ *tend to be an arbitrary choice, since all data points are indistinguishable whether they come from the first class or the second class.*

On the other hand, for the case where $\tau_1^2 > \tau_2^2$, *they showed that data from the first class tend to lie deterministically at the vertices of an* n_1*-simplex of edge length* $\sqrt{2}\tau_1\sqrt{p}$, *while data from the second class tend to lie deterministically at the vertices of an* n_2*-simplex of edge length* $\sqrt{2}\tau_2\sqrt{p}$ *and all pairwise angles are asymptotically orthogonal. Hence, data from the first class can asymptotically be explained only by the first* $(n_1 - 1)$ *sample eigenvectors in* $S_1 = \{\hat{\mathbf{u}}_1 \ldots, \hat{\mathbf{u}}_{n_1-1}\}$, *while data from the second class can be explained only by the rest of sample eigenvectors in* $S_2 = \{\hat{\mathbf{u}}_{n_1}, \ldots, \hat{\mathbf{u}}_{n-2}\}$. *Also, these eigenvectors can be arbitrarily chosen in each set.*

We will see that how assuming unequal tail eigenvalues affects data piling of independent test data.

Example 3 *We consider both classes have a common leading eigenvector, that is,* $\mathbf{u}_{(1),1} = \mathbf{u}_{(2),1} = \mathbf{u}_1$, *but this time we assume* $\tau_1^2 > \tau_2^2$ *instead of* $\tau_1^2 = \tau_2^2$ *in Example 1.*

In this case, both of $\hat{\mathbf{u}}_1$ *and* $\hat{\mathbf{u}}_{n_1}$ *are not strongly inconsistent with* $\mathbf{u}_1$, *while* $\hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_{n_1-1}$ *and* $\hat{\mathbf{u}}_{n_1+1}, \ldots, \hat{\mathbf{u}}_{n-2}$ *are strongly inconsistent with* $\mathbf{u}_1$. *Let* $\mathcal{D} = \{1, n_1\}$. *In Figure 3.2, independent test data* $\mathcal{Y}$ *projected onto* $\mathcal{S}_{n_1} =$

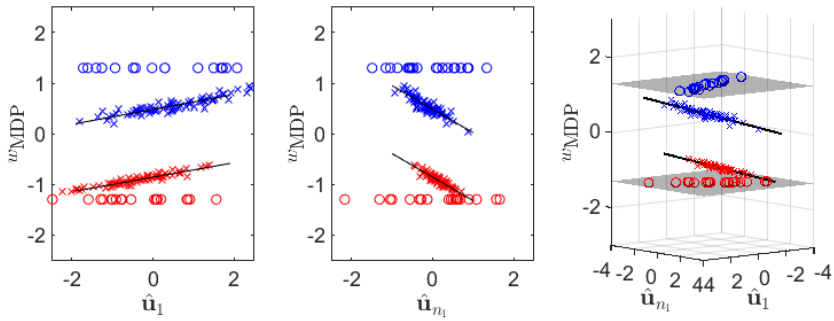

Figure 3.2 2-dimensional projections onto $\mathcal{S}_1 = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$ and $\mathcal{S}_{n_1} = \mathrm{span}(\hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ and 3-dimensional projections onto $\mathcal{S} = \mathrm{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\mathrm{MDP}})$ with $\mathcal{D} = \{1, n_1\}$ of training data $\mathcal{X}$ (class 1: blue circles, class 2: red circles) and independent test data $\mathcal{Y}$ (class 1: blue crosses, class 2: red crosses) under the model in Example 3.

$\mathrm{span}(\hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ *as well as those onto* $\mathcal{S}_1 = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$ *are also concentrated along parallel lines, one for each class. Thus, within the* 3*-dimensional subspace* $\mathcal{S} = \mathrm{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\mathrm{MDP}}) = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$*, the lines are parallel to each other. Also, they are asymptotically parallel to* $P_{\mathcal{S}}\mathbf{u}_1$*, which is the projection of the common leading eigenvector* $\mathbf{u}_1$ *onto* $\mathcal{S}$*. It implies that the variation of data along* $\mathbf{u}_1$ *is captured not only by* $\hat{\mathbf{u}}_1$ *but also by* $\hat{\mathbf{u}}_{n_1}$*.*

To understand this phenomenon, we focus on the geometric representation of HDLSS data. Jung et al. [2012] showed that in one class case, HDLSS data from strongly spiked covariance model can asymptotically be decomposed into random and deterministic parts; the random variation remains in $\mathrm{span}(\mathbf{u}_1)$*, while the deterministic structure (that is, the simplex described in Remark 1) remains in the orthogonal complement of* $\mathrm{span}(\mathbf{u}_1)$*. For sufficiently large* p*,* $\hat{\mathbf{u}}_1$ *explains the most important variation along* $\mathbf{u}_1$ *in the data from both*

classes, while $\hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_{n_1-1}$ *account for the deterministic simplex with edge length* $\sqrt{2}\tau_1\sqrt{p}$ *for data only from the first class. Then* $\hat{\mathbf{u}}_{n_1}$ *explains remaining variation along* $\mathbf{u}_1$ *in the data from both classes, which is smaller than the variance* τ_1^2 *of the first class but larger than the variance* τ_2^2 *of the second class. Lastly,* $\hat{\mathbf{u}}_{n_1+1}, \ldots, \hat{\mathbf{u}}_{n-2}$ *account for the deterministic simplex with edge length* $\sqrt{2}\tau_2\sqrt{p}$ *for data only from the second class. We emphasize that this result can be obtained with probability* 1*. Note that if* $\tau_1^2 = \tau_2^2$*, then only* $\hat{\mathbf{u}}_1$ *explains variation along* $\mathbf{u}_1$ *in the data, while the other sample eigenvectors explain the deterministic simplex for data from both classes.*

Example 4 *Two classes do not have a common leading eigenvector, that is,* $\mathbf{u}_{(1),1} \neq \mathbf{u}_{(2),1}$ *such that leading eigenvectors of each class form an angle of* $\pi/4$*, but this time we assume* $\tau_1^2 > \tau_2^2$ *instead of* $\tau_1^2 = \tau_2^2$ *in Example 2.*

As in the previous examples, $\hat{\mathbf{u}}_1$ *estimates the largest variation within the common leading eigenspace* $\mathcal{U}$ *from the data. However, in this example, the remaining variation may be either larger or smaller than* τ_1^2 *in contrast to the other examples. If the remaining variation within* $\mathcal{U}$ *is smaller than* τ_1^2*, then this variation is captured by* $\hat{\mathbf{u}}_{n_1}$*, while* $\hat{\mathbf{u}}_2$ *explains the deterministic simplex of data from the first class. Otherwise,* $\hat{\mathbf{u}}_2$ *captures the remaining variation, while* $\hat{\mathbf{u}}_{n_1}$ *explains the deterministic simplex of data from the first class. The other remaining sample eigenvectors are strongly inconsistent with* $\mathcal{U}$*.*

In Figure 3.3, independent test data $\mathcal{Y}$ *projected onto* $\mathcal{S}_1 = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$ *and* $\mathcal{S}_{n_1} = \mathrm{span}(\hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ *are concentrated along lines, but in both subspaces these lines are not parallel to each other. Also, independent test data* $\mathcal{Y}$ *projected onto* $\mathcal{S}_2 = \mathrm{span}(\hat{\mathbf{u}}_2, w_{\mathrm{MDP}})$ *are concentrated along lines, which is parallel to* w_{MDP}*, but these lines can completely overlap. However, within the 3-dimensional subspace* $\mathcal{S}_{1,n_1} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$*, there are two parallel 2-dimensional planes respectively including those lines. Similar to Ex-*

ample 2, $\mathcal{Y}_1$ *is distributed along the direction* $P_{\mathcal{S}_{1,n_1}}\mathbf{u}_{(1),1}$*, while* $\mathcal{Y}_2$ *is distributed along the direction* $P_{\mathcal{S}_{1,n_1}}\mathbf{u}_{(2),1}$*. Thus, these lines are asymptotically contained in 2-dimensional affine subspaces, that are parallel to* $P_{\mathcal{S}_{1,n_1}}\mathcal{U} = \mathrm{span}(P_{\mathcal{S}_{1,n_1}}\mathbf{u}_{(1),1}, P_{\mathcal{S}_{1,n_1}}\mathbf{u}_{(2),1})$.

Note that $\hat{\mathbf{u}}_2$ *instead of* $\hat{\mathbf{u}}_{n_1}$ *can capture the remaining variability within* $\mathcal{U}$ *depending on the true leading principal components scores of training data* $\mathcal{X}$*. Then 2-dimensional parallel affine subspaces can be observed in* $\mathcal{S}_{1,2} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, w_{\mathrm{MDP}})$ *instead of* $\mathcal{S}_{1,n_1} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$*. However, we can always observe 2-dimensional parallel affine subspaces in* $\mathcal{S} = \mathrm{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\mathrm{MDP}})$ *where* $\mathcal{D} = \{1, 2, n_1\}$.

The following proposition states that even if $\tau_1^2 > \tau_2^2$, projections of $\mathcal{Y}$ onto $\mathcal{S}$, which is a low-dimensional subspace of $\mathcal{S}_X$, are distributed along two parallel affine subspaces as p increases. However, in this case, $\mathcal{S}$ is not the subspace spanned by the first m eigenvectors of $\mathbf{S}_W$ and w_{MDP}.

Proposition 2 *Suppose Assumptions 1—5 hold. Also, assume* $\tau_1^2 > \tau_2^2$ *and* $m_1 = m_2 = 1$.

(i) If $m = 1$*, let* $\mathcal{S} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ *and* $L_k = \left\{\mathbf{u}_{1,\mathcal{S}}t + \nu_k w_{\mathrm{MDP}} + \bar{X}_{\mathcal{S}} : t \in \mathbb{R}\right\}$

(ii) If $m = 2$*, let* $\mathcal{S} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ *and* $L_k = \left\{\mathbf{U}_{1,\mathcal{S}}\mathbf{t} + \nu_k w_{\mathrm{MDP}} + \bar{X}_{\mathcal{S}} : \mathbf{t} \in \mathbb{R}^2\right\}$

for $k = 1, 2$ *where* $\nu_1 = \kappa_{\mathrm{MDP}}^{-1}(\eta_2(1 - \cos^2\varphi)\delta^2 - (\tau_1^2 - \tau_2^2)/n)$ *and* $\nu_2 = \kappa_{\mathrm{MDP}}^{-1}(-\eta_1(1 - \cos^2\varphi)\delta^2 - (\tau_1^2 - \tau_2^2)/n)$*. Then, for any independent observation* $Y \in \mathcal{Y}$ *and for any* $\epsilon > 0$*,*

$$\lim_{p\to\infty} \mathbb{P}\left(\inf_{a\in L_k} \|Y_{\mathcal{S}} - a\| > \epsilon | \pi(Y) = k\right) = 0$$

for $k = 1, 2$.

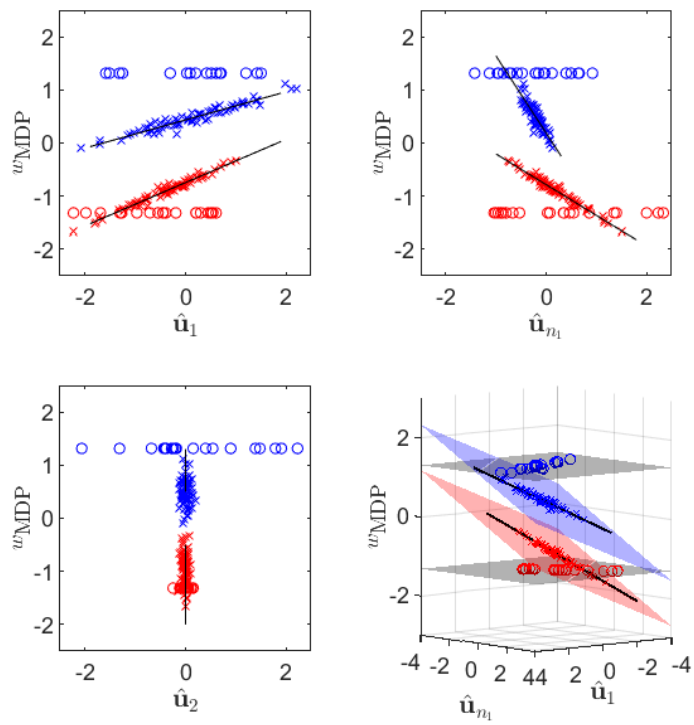

Figure 3.3 2-dimensional projections onto $\mathcal{S}_1 = \mathrm{span}(\hat{\mathbf{u}}_1, w_{\mathrm{MDP}})$, $\mathcal{S}_2 = \mathrm{span}(\hat{\mathbf{u}}_2, w_{\mathrm{MDP}})$ and $\mathcal{S}_{n_1} = \mathrm{span}(\hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ and 3-dimensional projections onto $\mathcal{S}_{1,n_1} = \mathrm{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_{n_1}, w_{\mathrm{MDP}})$ of training data $\mathcal{X}$ (class 1: blue circles, class 2: red circles) and independent test data $\mathcal{Y}$ (class 1: blue crosses, class 2: red crosses) under the model in Example 4.

3.2 Main Theorem

In this chapter, we extend Propositions 1 and 2 to the general cases where $m_1 \geq 1$ and $m_2 \geq 1$. We first characterize the signal subspace $\mathcal{S}$ for general cases where $m_1 \geq 1$ and $m_2 \geq 1$. For this, we investigate the asymptotic behavior of sample eigenvalues and eigenvectors of $\mathbf{S}_W$. For each $k = 1, 2$, denote the $n_k \times m_k$ matrix of the leading m_k principal component scores of the kth class as $\mathbf{W}_{(k)} = [\sigma_{(k),1} z_{(k),1}, \ldots, \sigma_{(k),m_k} z_{(k),m_k}]$. Also, denote the scaled covariance matrix of the leading m_k principal component scores of the kth class as

$$\mathbf{\Phi}_k = \mathbf{W}_{(k)}^\top (\mathbf{I}_{n_k} - \frac{1}{n_k} \mathbf{J}_{n_k}) \mathbf{W}_{(k)} \tag{3.3}$$

where $\mathbf{J}_{n_k}$ is the matrix of size $n_k \times n_k$ whose all entries are 1. Note that $\mathbf{\Phi}_1$, $\mathbf{\Phi}_2$ are symmetric positive definite matrices with probability 1. Let $\mathbf{W} = [\mathbf{R}_{(1)} \mathbf{W}_{(1)}^\top \ \mathbf{R}_{(2)} \mathbf{W}_{(2)}^\top]^\top$ and

$$\mathbf{\Phi} = \mathbf{W}^\top (\mathbf{I}_n - \mathbf{J}) \mathbf{W} \tag{3.4}$$

where $\mathbf{J} = \begin{pmatrix} \frac{1}{n_1} \mathbf{J}_{n_1} & \mathbf{O}_{n_1 \times n_2} \\ \mathbf{O}_{n_2 \times n_1} & \frac{1}{n_2} \mathbf{J}_{n_2} \end{pmatrix}$. Finally, let

$$\mathbf{\Phi}_{\tau_1, \tau_2} = \begin{pmatrix} \mathbf{\Phi}_1 + \tau_1^2 \mathbf{I}_{m_1} & \mathbf{\Phi}_1^{1/2} \mathbf{R}_{(1)}^\top \mathbf{R}_{(2)} \mathbf{\Phi}_2^{1/2} \\ \mathbf{\Phi}_2^{1/2} \mathbf{R}_{(2)}^\top \mathbf{R}_{(1)} \mathbf{\Phi}_1^{1/2} & \mathbf{\Phi}_2 + \tau_2^2 \mathbf{I}_{m_2} \end{pmatrix}. \tag{3.5}$$

Note that $\mathbf{\Phi}$ and $\mathbf{\Phi}_{\tau_1, \tau_2}$ are also symmetric positive definite matrices with probability 1. For any square matrix $\mathbf{M} \in \mathbb{R}^{l \times l}$ ($l \in \mathbb{N}$), let $\phi_i(\mathbf{M})$ and $v_i(\mathbf{M})$ denote the ith largest eigenvalue of $\mathbf{M}$ and its corresponding eigenvector, respectively. The following lemma shows asymptotic behavior of sample eigenvalues of $\mathbf{S}_W$. Throughout, we assume $\tau_1^2 \geq \tau_2^2$.

Lemma 3 *Suppose Assumptions 1—5 hold. Then, the following hold as* $p \to \infty$.

(i) If $\tau_1^2 = \tau_2^2 =: \tau^2$, *then conditional to* $\mathbf{W}_{(1)}$ *and* $\mathbf{W}_{(2)}$,

$$p^{-1}\hat{\lambda}_i \xrightarrow{P} \begin{cases} \phi_i(\mathbf{\Phi}) + \tau^2, & 1 \le i \le m, \\ \tau^2, & m+1 \le i \le n-2. \end{cases}$$

(ii) If $\tau_1^2 > \tau_2^2$, *then conditional to* $\mathbf{W}_{(1)}$ *and* $\mathbf{W}_{(2)}$,

$$p^{-1}\hat{\lambda}_i \xrightarrow{P} \begin{cases} \phi_i(\mathbf{\Phi}_{\tau_1,\tau_2}), & 1 \le i \le k_0, \\ \tau_1^2, & k_0+1 \le i \le k_0 + (n_1 - m_1 - 1), \\ \phi_{i-(n_1-m_1-1)}(\mathbf{\Phi}_{\tau_1,\tau_2}), & k_0 + (n_1 - m_1) \le i \le n_1 + m_2 - 1, \\ \tau_2^2, & n_1 + m_2 \le i \le n-2, \end{cases}$$

where k_0 $(m_1 \le k_0 \le m_1 + m_2)$ *is an integer which satisfies* $\phi_{k_0}(\mathbf{\Phi}_{\tau_1,\tau_2}) \ge \tau_1^2 \ge \phi_{k_0+1}(\mathbf{\Phi}_{\tau_1,\tau_2})$ *if we denote* $\phi_{m_1+m_2+1}(\mathbf{\Phi}_{\tau_1,\tau_2}) = 0$.

Remark 2 *(i) If* $\tau_1^2 = \tau_2^2 =: \tau^2$, *then*

$$\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2}) = \begin{cases} \phi_i(\mathbf{\Phi}) + \tau^2, & 1 \le i \le m, \\ \tau^2, & m+1 \le i \le m_1 + m_2. \end{cases}.$$

Thus, Lemma 3 (i) can be seen as a special case of Lemma 3 (ii).

(ii) If $\tau_1^2 > \tau_2^2$ *and* $m = m_1$, *then* k_0 *in Lemma 3 (ii) is* m_1 *with probability* 1 *by Weyl's inequality.*

Lemma 3 shows that the asymptotic behavior of sample eigenvalues of $\mathbf{S}_W$ is quite different depending on whether both covariance matrices have equal tail eigenvalues or unequal tail eigenvalues. If $\tau_1^2 = \tau_2^2$, then the first m sample eigenvalues explain true leading principal component scores of both classes, while the other sample eigenvalues do not. In contrast, if $\tau_1^2 \neq \tau_2^2$, we observe a counter-intuitive phenomenon that some non-leading sample eigenvalues can

explain true leading principal component scores instead of some leading sample eigenvalues.

The following lemma gives the limiting angle between $\hat{\mathbf{u}}_i$ and the common leading eigenspace $\mathcal{U}$.

Lemma 4 *Suppose Assumptions 1—5 hold. Then, the following hold as* $p \to \infty$.

(i) If $\tau_1^2 = \tau_2^2 =: \tau^2$, *then conditional to* $\mathbf{W}_{(1)}$ *and* $\mathbf{W}_{(2)}$,

$$\cos\left(Angle(\hat{\mathbf{u}}_i, \mathcal{U})\right) \xrightarrow{P} \begin{cases} C_i, & 1 \le i \le m, \\ 0, & m+1 \le i \le n-2 \end{cases}$$

where

$$C_i = \sqrt{\frac{\phi_i(\mathbf{\Phi})}{\phi_i(\mathbf{\Phi}) + \tau^2}} > 0. \tag{3.6}$$

(ii) If $\tau_1^2 > \tau_2^2$ *and* $m > m_1$, *then conditional to* $\mathbf{W}_{(1)}$ *and* $\mathbf{W}_{(2)}$,

$$\cos\left(Angle(\hat{\mathbf{u}}_i, \mathcal{U})\right)$$
$$\xrightarrow{P} \begin{cases} D_i, & 1 \le i \le k_0, \\ 0, & k_0 + 1 \le i \le k_0 + (n_1 - m_1 - 1), \\ D_{i-(n_1-m_1-1)}, & k_0 + (n_1 - m_1) \le i \le n_1 + m_2 - 1, \\ 0, & n_1 + m_2 \le i \le n - 2 \end{cases}$$

where k_0 *is defined in Lemma 3 (ii) and*

$$D_i = \sqrt{\frac{\| \sum_{k=1}^{2} \mathbf{R}_{(k)} \mathbf{\Phi}_k^{1/2} \tilde{v}_{ik}(\mathbf{\Phi}_{\tau_1,\tau_2}) \|^2}{\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})}} > 0. \tag{3.7}$$

Here, $v_i(\mathbf{\Phi}_{\tau_1,\tau_2}) = (\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2})^\top, \tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2})^\top)^\top$ *with* $\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2}) \in \mathbb{R}^{m_1}$ *and* $\tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2}) \in \mathbb{R}^{m_2}$.

We define an index set $\mathcal{D} \subset \{1, \ldots, n-2\}$ for general cases where $m_1 \geq 1$ and $m_2 \geq 1$. Let $i \in \mathcal{D}$ if and only if there exists $\epsilon > 0$ such that $\lim_{p\to\infty} \mathbb{P}(\cos(\text{Angle}(\hat{\mathbf{u}}_i, \mathcal{U})) > \epsilon) > 0$. In contrast, $i \notin \mathcal{D}$ if and only if $\hat{\mathbf{u}}_i$ is strongly inconsistent with the common leading eigenspace $\mathcal{U}$ in the sense that $\text{Angle}(\hat{\mathbf{u}}_i, \mathcal{U}) \xrightarrow{P} \pi/2$ as $p \to \infty$. In other words, $\hat{\mathbf{u}}_i$ with $i \notin \mathcal{D}$ is a noisy direction which does not capture important variability within the common leading eigenspace $\mathcal{U}$, while $\hat{\mathbf{u}}_i$ with $i \in \mathcal{D}$ may explain important variability.

Note that if we further assume $m = m_1$ (that is, $\mathcal{U} = \mathcal{U}_{(1)}$) in Lemma 4 (ii), then $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m$ explain the most important variation within $\mathcal{U}$, while $\hat{\mathbf{u}}_{n_1}, \ldots, \hat{\mathbf{u}}_{n_1+m_2-1}$ explain the remaining variation within $\mathcal{U}$. The other sample eigenvectors do not explain the variability. Hence, $(m_1 + m_2)$ sample eigenvectors are needed to explain the variation within $\mathcal{U}$ (See Example 3).

If $m > m_1$, then for given training data $\mathcal{X}$, $(m_1 + m_2)$ sample eigenvectors, which are $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{k_0}$ and $\hat{\mathbf{u}}_{k_0+(n_1-m_1)}, \ldots, \hat{\mathbf{u}}_{n_1+m_2-1}$, explain the variation within $\mathcal{U}$. However, k_0 $(m_1 \leq k_0 \leq m_1 + m_2)$ is a random number depending on true leading principal component scores $\mathbf{W}_{(1)}$ and $\mathbf{W}_{(2)}$. This fact implies that, in general, if $\mathbb{P}(k_0 = i) > 0$ for all $m_1 \leq k_0 \leq m_1 + m_2$, then $(m_1 + 2m_2)$ sample eigenvectors, which are $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{m_1+m_2}$ and $\hat{\mathbf{u}}_{n_1}, \ldots, \hat{\mathbf{u}}_{n_1+m_2-1}$, are all needed to explain the variation within $\mathcal{U}$ (See Example 4).

From Lemma 4, we can characterize $\mathcal{D}$ for general cases where $m_1 \geq 1$ and $m_2 \geq 1$. We summarize $\mathcal{D}$ in Table 3.1 for each case. In general, we define the signal subspace $\mathcal{S}$ as

$$\mathcal{S} = \text{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\text{MDP}}), \tag{3.8}$$

which is obtained by removing the noisy directions in the sample space $\mathcal{S}_X$.

We now confirm that projections of $\mathcal{Y}$ onto $\mathcal{S}$ in (3.8), which is a low-dimensional subspace of the sample space $\mathcal{S}_X$, are distributed along parallel

Condition	$\mathcal{D}$	$\lvert\mathcal{D}\rvert$
$\tau_1^2 = \tau_2^2$	$\{1, \ldots, m\}$	m
$\tau_1^2 > \tau_2^2$ $m = m_1$	$\{1, \ldots, m_1, n_1, \ldots, n_1 + m_2 - 1\}$	$m_1 + m_2$
$\tau_1^2 > \tau_2^2$ $m > m_1$	$\{1, \ldots, m_1 + m_2, n_1, \ldots, n_1 + m_2 - 1\}$	$m_1 + 2m_2$

Table 3.1 The index set $\mathcal{D}$ for each case.

affine subspaces, one for each class, and that those affine subspaces do not coincide. Recall that κ_{MDP} is the training data piling distance defined in (3.2).

Theorem 5 *Suppose Assumptions 1—5 hold. Let* $\mathcal{S} = \text{span}(\{\hat{\mathbf{u}}_i\}_{i\in\mathcal{D}}, w_{\text{MDP}})$ *with* $\mathcal{D}$ *be given in Table 3.1 for each case. Also, let*

$$L_k = \left\{\mathbf{U}_{1,\mathcal{S}}\mathbf{t} + \nu_k w_{\text{MDP}} + \bar{X}_{\mathcal{S}} : \mathbf{t} \in \mathbb{R}^m\right\}$$

for $k = 1, 2$ *where*

$$\nu_1 = \kappa_{\text{MDP}}^{-1}\left(\eta_2(1 - \cos^2\varphi)\delta^2 - (\tau_1^2 - \tau_2^2)/n\right)$$

and

$$\nu_2 = \kappa_{\text{MDP}}^{-1}\left(-\eta_1(1 - \cos^2\varphi)\delta^2 - (\tau_1^2 - \tau_2^2)/n\right)$$

Then for any independent observation $Y \in \mathcal{Y}$ *and for any* $\epsilon > 0$,

$$\lim_{p\to\infty} \mathbb{P}\left(\inf_{a\in L_k} \|Y_{\mathcal{S}} - a\| > \epsilon | \pi(Y) = k\right) = 0$$

for $k = 1, 2$.

Remark 3 *Write the projections of* $\mathbf{u}_{(k),i}$ $(k = 1, 2)$ *onto a subspace* $\mathcal{S}$ *of* $\mathbb{R}^p$ *as* $\mathbf{u}_{(k),i,\mathcal{S}} = P_{\mathcal{S}}\mathbf{u}_{(k),i}$ *and* $\mathbf{U}_{(k),1,\mathcal{S}} = [\mathbf{u}_{(k),1,\mathcal{S}}, \ldots, \mathbf{u}_{(k),m_k,\mathcal{S}}]$ *for* $k = 1, 2$. *Then*

projections of $\mathcal{Y}_1$ *are distributed along an* m_1*-dimensional affine subspace* L_1'*, which is parallel to* $\mathrm{span}(\mathbf{U}_{(1),1,\mathcal{S}})$*, while projections of* $\mathcal{Y}_2$ *are distributed along an* m_2*-dimensional affine subspace* L_2'*, which is parallel to* $\mathrm{span}(\mathbf{U}_{(2),1,\mathcal{S}})$*. For each* $k = 1, 2$*, the* m*-dimensional affine subspace* L_k *contains* L_k'*.*

Theorem 5 tells that independent test data are asymptotically distributed along parallel m-dimensional affine subspaces L_1 and L_2 in $\mathcal{S}$. It implies that if we find a direction $w \in \mathcal{S}$ such that w is asymptotically orthogonal to L_1 and L_2, then $P_w\mathcal{Y}$ yields the second data piling and in turn achieves perfect classification of independent test data. Since $\dim \mathcal{S} - m \geq 1$, there always exists a direction $w \in \mathcal{S}$ which yields second data piling. Among second data piling directions, we will find a second maximal data piling direction, which provides asymptotic maximal distance between the two piles of independent test data among second data piling directions.

Chapter 4

Estimation of Second Maximal Data Piling Direction

In this chapter, we propose novel algorithms to estimate a second maximal data piling direction. Let $\mathbf{V} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n-2}, w_{\mathrm{MDP}}]$, which collects an orthonormal basis of the sample space $\mathcal{S}_X$. Also, in this chapter, we assume that an independent test dataset $\mathcal{Y}$ is available to us (It is possible by splitting the original training dataset $\mathcal{X}$ into the new training dataset $\mathcal{X}$ and the test dataset $\mathcal{Y}$). Denote the horizontally concatenated data matrix of the given independent test dataset $\mathcal{Y}$ by

$$\mathbf{Y} = [Y_{11}, \ldots, Y_{1n_1^*}, Y_{21}, \ldots, Y_{2n_2^*}].$$

The $p \times n^*$ data matrix $\mathbf{Y}$ consists of the $n^* := n_1^* + n_2^*$ observations independent to $\mathcal{X}$ and arranged so that $\pi(Y_{kj}) = k$ for any k, j. We assume that n_k^* is fixed and $n_k^* > m_k$ for $k = 1, 2$. Write class-wise sample mean vectors $\bar{Y}_k = n_k^{*-1} \sum_{j=1}^{n_k^*} Y_{kj}$. We define the within-scatter matrix of $\mathcal{Y}$ as

$$\mathbf{S}_W^* = (\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top,$$

where $\bar{\mathbf{Y}} = [\bar{\mathbf{Y}}_1\ \bar{\mathbf{Y}}_2]$ and $\bar{\mathbf{Y}}_k = \bar{Y}_k \mathbf{1}_{n_k^*}^\top$ for $k = 1, 2$.

We will find a sequence of directions $\{w\} \in \mathfrak{W}_X$ which yields second data piling for given independent test dataset $\mathcal{Y}$. The condition that $p^{-1/2} w^\top (Y - Y') \xrightarrow{P} 0$ as $p \to \infty$ for any $Y, Y' \in \mathcal{Y}$ with $\pi(Y) = \pi(Y')$ is equivalent to the condition that

$$\frac{1}{p} w^\top \mathbf{S}_W^* w \xrightarrow{P} 0$$

as $p \to \infty$. Thus, we define the collection of sequences of second data piling directions for $\mathcal{Y}$ as

$$\bar{\mathcal{A}} = \left\{ \{w\} \in \mathfrak{W}_X : \frac{1}{p} w^\top \mathbf{S}_W^* w \xrightarrow{P} 0 \text{ as } p \to \infty \right\}.$$

For any $\{w\} \in \mathfrak{W}_X$, we can write $w = \mathbf{V}\mathbf{a}$ for some $\mathbf{a} = (a_1, \ldots, a_{n-2}, a_{\text{MDP}})^\top \in \mathbb{R}^{n-1}$. Without loss of generality, we assume $a_{\text{MDP}} \geq 0$ for all p. For $\{w\} \in \bar{\mathcal{A}}$, we can write

$$\frac{1}{p} w^\top \mathbf{S}_W^* w = \frac{1}{p} \mathbf{a}^\top \mathbf{V}^\top \mathbf{S}_W^* \mathbf{V} \mathbf{a} = \mathbf{a}^\top \left(\frac{1}{p} \mathbf{V}^\top \mathbf{S}_W^* \mathbf{V} \right) \mathbf{a}. \tag{4.1}$$

Note that the $(n-1) \times (n-1)$ matrix $p^{-1} \mathbf{V}^\top \mathbf{S}_W^* \mathbf{V}$ can be understood as the scatter of the independent test data $\mathcal{Y}$ projected onto the sample space $\mathcal{S}_X$. Theorem 6 shows that independent test data $\mathcal{Y}$ are asymptotically supported on a m-dimensional subspace in $\mathcal{S}_X$.

Theorem 6 *$p^{-1} \mathbf{V}^\top \mathbf{S}_W^* \mathbf{V}$ converges to a rank m matrix in probability as $p \to \infty$.*

We write an eigen-decomposition of $p^{-1} \mathbf{V}^\top \mathbf{S}_W^* \mathbf{V} = \hat{\mathbf{Q}} \mathbf{H} \hat{\mathbf{Q}}^\top$, where $\mathbf{H} = \text{Diag}(h_1, \ldots, h_{n-1})$ arranged in descending order, and $\hat{\mathbf{Q}} = [\hat{\mathbf{Q}}_1\ \hat{\mathbf{Q}}_2]$ with $\hat{\mathbf{Q}}_1 = [\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_m]$ and $\hat{\mathbf{Q}}_2 = [\hat{\mathbf{q}}_{m+1}, \ldots, \hat{\mathbf{q}}_{n-1}]$. Meanwhile, $\mathbf{a}$ can be written as $\mathbf{a} = \hat{\mathbf{Q}} \boldsymbol{\iota} = \sum_{i=1}^{n-1} \iota_i \hat{\mathbf{q}}_i$ for some sequence of $\boldsymbol{\iota} = (\iota_1, \ldots, \iota_{n-1})^\top$. Since

$$\frac{1}{p} w^\top \mathbf{S}_W^* w = \mathbf{a}^\top \left(\frac{1}{p} \mathbf{V}^\top \mathbf{S}_W^* \mathbf{V} \right) \mathbf{a} = \sum_{i=1}^{n-1} h_i \iota_i^2 = \sum_{i=1}^{m} h_i \iota_i^2 + o_p(1)$$

by Theorem 6 and (4.1), $\{w\} \in \bar{\mathcal{A}}$ if and only if $\iota_1, \ldots, \iota_m \xrightarrow{P} 0$ as $p \to \infty$. In other words, for any given $\{w\} \in \bar{\mathcal{A}}$, there exists a sequence of directions $\{v\}$ such that $v \in \text{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ for all p and $\|w - v\| \xrightarrow{P} 0$ as $p \to \infty$. This fact plays a crucial role in our next observation: $\{w\} \in \bar{\mathcal{A}}$ is indeed asymptotically orthogonal to the common leading eigenspace $\mathcal{U}$.

Theorem 7 *Suppose Assumptions 1—5 hold. For any given* $\{w\} \in \bar{\mathcal{A}}$, $w^\top \mathbf{u}_j \xrightarrow{P} 0$ *for* $1 \le j \le m$.

Furthermore, $\{w\} \in \bar{\mathcal{A}}$ can also achieve perfect classification of any independent observation Y, which is independent to both of $\mathcal{X}$ and $\mathcal{Y}$. Theorem 8 confirms that we can achieve perfect classification if we choose $\mathbf{a}$ so that for $w = \mathbf{V}\mathbf{a}$, $\{w\} \in \bar{\mathcal{A}}$ and $\lim_{p\to\infty} a_{\text{MDP}} > 0$.

Theorem 8 *Suppose Assumptions 1—5 hold. For any given* $\{w\} \in \bar{\mathcal{A}}$, *write*

$$w = \mathbf{V}\mathbf{a} = \sum_{k=1}^{n-2} a_k \hat{\mathbf{u}}_k + a_{\text{MDP}} w_{\text{MDP}}$$

with $\mathbf{a} = (a_1, \ldots, a_{n-2}, a_{\text{MDP}})^\top$ *and assume* $a_{\text{MDP}} \xrightarrow{P} \psi_{\text{MDP}}$ *as* $p \to \infty$. *Then for any independent observation* Y, *which is independent to both of* $\mathcal{X}$ *and* $\mathcal{Y}$,

$$\frac{1}{\sqrt{p}} w^\top (Y - \bar{X}) \xrightarrow{P} \begin{cases} \frac{\psi_{\text{MDP}}}{\kappa}(\eta_2(1 - \cos^2\varphi)\delta^2 - (\tau_1^2 - \tau_2^2)/n), & \pi(Y) = 1, \\ \frac{\psi_{\text{MDP}}}{\kappa}(-\eta_1(1 - \cos^2\varphi)\delta^2 - (\tau_1^2 - \tau_2^2)/n), & \pi(Y) = 2, \end{cases}$$

as $p \to \infty$, *where* κ *is the probability limit of* κ_{MDP} *defined in (3.2).*

Theorem 8 also shows that an asymptotic distance between the two piles of independent test data, which are independent to both of $\mathcal{X}$ and $\mathcal{Y}$, can be maximized if $\{w\} \in \bar{\mathcal{A}}$ with $w = \mathbf{V}\mathbf{a}$ has a maximal limit of a_{MDP}. Theorem 9 confirms that a projection of w_{MDP} onto $\text{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ is an estimate of a second *maximal* data piling direction. Recall that $\mathbf{V}\hat{\mathbf{Q}}_2$ is obtainable by using $\mathcal{X}$ and

$\mathcal{Y}$, and the dimension of $\text{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ is $n-m-1$. It implies that a second maximal data piling direction can be obtained by projecting w_{MDP} onto the nullspace of the common leading eigenspace $\mathcal{U}$.

Theorem 9 *Suppose Assumptions 1—5 hold. Write* $\mathbf{e}_{\text{MDP}} = (\mathbf{0}_{n-2}^\top,\ 1)^\top$ *so that* $w_{\text{MDP}} = \mathbf{V}\mathbf{e}_{\text{MDP}}$. *Also, let* $\{w_{\text{SMDP}}\}$ *be a sequence of directions such that* $w_{\text{SMDP}} = \mathbf{V}\mathbf{a}_{\text{SMDP}}$ *where*

$$\mathbf{a}_{\text{SMDP}} = \frac{P_{\text{span}(\hat{\mathbf{Q}}_2)}\mathbf{e}_{\text{MDP}}}{\|P_{\text{span}(\hat{\mathbf{Q}}_2)}\mathbf{e}_{\text{MDP}}\|} = \frac{\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top\mathbf{e}_{\text{MDP}}}{\|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top\mathbf{e}_{\text{MDP}}\|} \in \mathbb{R}^{n-1}.$$

Then $\{w\} \in \bar{\mathcal{A}}$ *is a sequence of second maximal data piling directions if and only if* $\|w - w_{\text{SMDP}}\| \xrightarrow{P} 0$ *as* $p \to \infty$.

We have shown that a second maximal data piling direction in the sample space $\mathcal{S}_X$ can be obtained with a help of independent test data. As such, we randomly split $\mathcal{X}_k$, which is the original training dataset of the kth class, into training dataset $\mathcal{X}_{k,tr}$ and test dataset $\mathcal{X}_{k,te}$ so that the sample size of test data of kth class $n_{k,te}$ is larger than m_k for $k = 1, 2$. Then we can find a second maximal data piling direction in the sample space of $\mathcal{X}_{tr} = \mathcal{X}_{1,tr} \cup \mathcal{X}_{2,tr}$ with a help of $\mathcal{X}_{te} = \mathcal{X}_{1,te} \cup \mathcal{X}_{2,te}$.

The fact that the sample size of HDLSS data is very small implies that classification using one data split may be unreliable (albeit theoretically true). In order to resolve this concern, we repeat the above procedure several times and set a final estimate of a second maximal data piling direction as the average of estimates of a second maximal data piling direction obtained from each repetition. A detailed algorithm is given in Algorithm 1. In practice, we should estimate m_1, m_2 and m, which are the true numbers of leading eigenvalues of $\mathbf{\Sigma}_{(1)}$, $\mathbf{\Sigma}_{(2)}$ and $\mathbf{\Sigma}_{(0)} = \pi_1\mathbf{\Sigma}_{(1)} + \pi_2\mathbf{\Sigma}_{(2)}$ for Algorithm 1. Estimating those numbers is feasible by Kritchman and Nadler [2008], Leek [2010], Passemier and Yao [2014] and Jung et al. [2018].

Algorithm 1 Second Maximal Data Piling (SMDP) algorithm (Type I)

Require: Original training data matrix of the kth class $\mathbf{X}_k$ for $k = 1, 2$.

Require: The number of repetitions K, estimated m_1, m_2 and m

1: **for** $j = 1, \ldots, K$ **do**

2: Randomly split $\mathbf{X}_k$ into $\begin{cases} \mathbf{X}_{k,tr} = [X_{k,1}, \ldots, X_{k,n_{k,tr}}] \\ \mathbf{X}_{k,te} = [X^*_{k,1}, \ldots, X^*_{k,n_{k,te}}] \end{cases}$ so that $n_{k,te} > m_k$

3: Set $n_{tr} = n_{1,tr} + n_{2,tr}$, $\bar{X}_{k,tr} = n_k^{-1}\mathbf{X}_{k,tr}\mathbf{1}_{n_{k,tr}}$ and $\bar{\mathbf{X}}_{k,tr} = \bar{X}_{k,tr}\mathbf{1}^\top_{n_{k,tr}}$

4: Set $\mathbf{S}_{tr} = \sum_{k=1}^{2}(\mathbf{X}_{k,tr} - \bar{\mathbf{X}}_{k,tr})(\mathbf{X}_{k,tr} - \bar{\mathbf{X}}_{k,tr})^\top$ and $\mathbf{d}_{tr} = \bar{X}_{1,tr} - \bar{X}_{2,tr}$

5: Write an eigen-decomposition of $\mathbf{S}_{tr}$ by $\mathbf{S}_{tr} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top = \hat{\mathbf{U}}_1\hat{\mathbf{\Lambda}}_1\hat{\mathbf{U}}_1^\top$ where $\hat{\mathbf{\Lambda}} = \mathrm{Diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{n_{tr}-2}, 0, \ldots, 0)$ arranged in descending order, and $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_1\ \hat{\mathbf{U}}_2]$ with $\hat{\mathbf{U}}_1 = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n_{tr}-2}]$, $\hat{\mathbf{U}}_2 = [\hat{\mathbf{u}}_{n_{tr}-1}, \ldots, \hat{\mathbf{u}}_p]$

6: Set $w_{\mathrm{MDP}} = \|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^\top\mathbf{d}_{tr}\|^{-1}\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^\top\mathbf{d}_{tr}$ and $\kappa_{\mathrm{MDP}} = p^{-1/2}\|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^\top\mathbf{d}_{tr}\|$

7: Set $\mathbf{V} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n_{tr}-2}, w_{\mathrm{MDP}}]$

8: Set $\bar{X}_{k,te} = n_{k,te}^{-1}\mathbf{X}_{k,te}\mathbf{1}_{n_{k,te}}$ and $\bar{\mathbf{X}}_{k,te} = \bar{X}_{k,te}\mathbf{1}^\top_{n_{k,te}}$

9: Set $\mathbf{S}_{te} = \sum_{k=1}^{2}(\mathbf{X}_{k,te} - \bar{\mathbf{X}}_{k,te})(\mathbf{X}_{k,te} - \bar{\mathbf{X}}_{k,te})^\top$

10: Write an eigen-decomposition of $p^{-1}\mathbf{V}^\top\mathbf{S}_{te}\mathbf{V}$ by $p^{-1}\mathbf{V}^\top\mathbf{S}_{te}\mathbf{V} = \hat{\mathbf{Q}}\mathbf{H}\hat{\mathbf{Q}}^\top$ where $\mathbf{H} = \mathrm{Diag}(h_1, \ldots, h_{n_{tr}-1})$ arranged in descending order, and $\hat{\mathbf{Q}} = [\hat{\mathbf{Q}}_1\ \hat{\mathbf{Q}}_2]$ with $\hat{\mathbf{Q}}_1 = [\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_m]$ and $\hat{\mathbf{Q}}_2 = [\hat{\mathbf{q}}_{m+1}, \ldots, \hat{\mathbf{q}}_{n_{tr}-1}]$

11: Set $\mathbf{a}_{j,\mathrm{SMDP}} = \|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top\mathbf{e}_{\mathrm{MDP}}\|^{-1}\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top\mathbf{e}_{\mathrm{MDP}}$ where $\mathbf{e}_{\mathrm{MDP}} = (\mathbf{0}^\top_{n-2}, 1)^\top$

12: Set $w_{j,\mathrm{SMDP}} = \mathbf{V}\mathbf{a}_{j,\mathrm{SMDP}}$

13: Set $\bar{X}_{j,\mathrm{SMDP}} = n_{tr}^{-1}(n_{1,tr}\bar{X}_{1,tr} + n_{2,tr}\bar{X}_{2,tr})$

14: Set $\hat{\alpha}_k = -p^{-1}\sum_{k=m_k+1}^{n_{k,tr}-1}\hat{\lambda}_{(k),l}$ where $\hat{\lambda}_{(k),l}$ is lth largest eigenvalue of

$$\mathbf{S}_{k,tr} = (\mathbf{X}_{k,tr} - \bar{\mathbf{X}}_{k,tr})(\mathbf{X}_{k,tr} - \bar{\mathbf{X}}_{k,tr})^\top$$

15: Set $g_{j,\mathrm{SMDP}} = (n_{tr}\kappa_{\mathrm{MDP}})^{-1}(\mathbf{e}^\top_{\mathrm{MDP}}\mathbf{a}_{j,\mathrm{SMDP}})(\hat{\alpha}_1 - \hat{\alpha}_2)$

16: **end for**

17: Set $w_{\mathrm{SMDP}} = K^{-1}\sum_{j=1}^{K} w_{j,\mathrm{SMDP}}$

18: Set $\bar{X}_{\mathrm{SMDP}} = K^{-1}\sum_{j=1}^{K} w^\top_{j,\mathrm{SMDP}}\bar{X}_{j,\mathrm{SMDP}}$ and $g_{\mathrm{SMDP}} = K^{-1}\sum_{j=1}^{K} g_{j,\mathrm{SMDP}}$

19: Use the following classification rule:

$$\phi_{\text{SMDP-I}}(Y; \mathcal{X}) = \begin{cases} 1, & p^{-1/2}(w^\top_{\mathrm{SMDP}}Y - \bar{X}_{\mathrm{SMDP}}) - g_{\mathrm{SMDP}} \geq 0, \\ 2, & p^{-1/2}(w^\top_{\mathrm{SMDP}}Y - \bar{X}_{\mathrm{SMDP}}) - g_{\mathrm{SMDP}} < 0. \end{cases} \tag{4.2}$$

Algorithm 2 Second Maximal Data Piling (SMDP) algorithm (Type II)

Require: Original training data matrix of the kth class $\mathbf{X}_k$ for $k = 1, 2$.

Require: The number of repetitions K, estimated m

1: **for** $j = 1, \ldots, K$ **do**

2: Randomly split $\mathbf{X}_k$ into $\begin{cases} \mathbf{X}_{k,tr} = [X_{k,1}, \ldots, X_{k,n_{k,tr}}] \\ \mathbf{X}_{k,te} = [X^*_{k,1}, \ldots, X^*_{k,n_{k,te}}] \end{cases}$ so that $n_{k,te} > m_k$

3: Set $n_{tr} = n_{1,tr} + n_{2,tr}$, $\bar{X}_{k,tr} = n_k^{-1}\mathbf{X}_{k,tr}\mathbf{1}_{n_{k,tr}}$ and $\bar{\mathbf{X}}_{k,tr} = \bar{X}_{k,tr}\mathbf{1}^\top_{n_{k,tr}}$

4: Set $\mathbf{S}_{tr} = \sum_{k=1}^{2}(\mathbf{X}_{k,tr} - \bar{\mathbf{X}}_{k,tr})(\mathbf{X}_{k,tr} - \bar{\mathbf{X}}_{k,tr})^\top$ and $\mathbf{d}_{tr} = \bar{X}_{1,tr} - \bar{X}_{2,tr}$

5: Write an eigen-decomposition of $\mathbf{S}_{tr}$ by $\mathbf{S}_{tr} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^\top = \hat{\mathbf{U}}_1\hat{\mathbf{\Lambda}}_1\hat{\mathbf{U}}_1^\top$ where $\hat{\mathbf{\Lambda}} = \mathrm{Diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{n_{tr}-2}, 0, \ldots, 0)$ arranged in descending order, and $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_1\ \hat{\mathbf{U}}_2]$ with $\hat{\mathbf{U}}_1 = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n_{tr}-2}]$, $\hat{\mathbf{U}}_2 = [\hat{\mathbf{u}}_{n_{tr}-1}, \ldots, \hat{\mathbf{u}}_p]$

6: Set $w_{\mathrm{MDP}} = \|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^\top\mathbf{d}_{tr}\|^{-1}\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^\top\mathbf{d}_{tr}$

7: Set $\mathbf{V} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n_{tr}-2}, w_{\mathrm{MDP}}]$

8: Set $\bar{X}_{k,te} = n_{k,te}^{-1}\mathbf{X}_{k,te}\mathbf{1}_{n_{k,te}}$ and $\bar{\mathbf{X}}_{k,te} = \bar{X}_{k,te}\mathbf{1}^\top_{n_{k,te}}$

9: Set $\mathbf{S}_{te} = \sum_{k=1}^{2}(\mathbf{X}_{k,te} - \bar{\mathbf{X}}_{k,te})(\mathbf{X}_{k,te} - \bar{\mathbf{X}}_{k,te})^\top$

10: Write an eigen-decomposition of $p^{-1}\mathbf{V}^\top\mathbf{S}_{te}\mathbf{V}$ by $p^{-1}\mathbf{V}^\top\mathbf{S}_{te}\mathbf{V} = \hat{\mathbf{Q}}\mathbf{H}\hat{\mathbf{Q}}^\top$ where $\mathbf{H} = \mathrm{Diag}(h_1, \ldots, h_{n_{tr}-1})$ arranged in descending order, and $\hat{\mathbf{Q}} = [\hat{\mathbf{Q}}_1\ \hat{\mathbf{Q}}_2]$ with $\hat{\mathbf{Q}}_1 = [\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_m]$ and $\hat{\mathbf{Q}}_2 = [\hat{\mathbf{q}}_{m+1}, \ldots, \hat{\mathbf{q}}_{n_{tr}-1}]$

11: Set $\mathbf{a}_{j,\mathrm{SMDP}} = \|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top\mathbf{e}_{\mathrm{MDP}}\|^{-1}\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top\mathbf{e}_{\mathrm{MDP}}$ where $\mathbf{e}_{\mathrm{MDP}} = (\mathbf{0}^\top_{n-2}, 1)^\top$

12: Set $w_{j,\mathrm{SMDP}} = \mathbf{V}\mathbf{a}_{j,\mathrm{SMDP}}$

13: Apply Linear Discriminant Analysis to $p^{-1/2}w^\top_{j,\mathrm{SMDP}}\mathbf{X}_{te}$ where $\mathbf{X}_{te} = [\mathbf{X}_{1,te}\ \mathbf{X}_{2,te}]$ and achieve a classification threshold b_j.

14: **end for**

15: Set $w_{\mathrm{SMDP}} = K^{-1}\sum_{j=1}^{K} w_{j,\mathrm{SMDP}}$

16: Set $b_{\mathrm{SMDP}} = K^{-1}\sum_{j=1}^{K} b_j$

17: Use the following classification rule:

$$\phi_{\text{SMDP-II}}(Y; \mathcal{X}) = \begin{cases} 1, & p^{-1/2}w^\top_{\mathrm{SMDP}}Y \geq b_{\mathrm{SMDP}}, \\ 2, & p^{-1/2}w^\top_{\mathrm{SMDP}}Y < b_{\mathrm{SMDP}}. \end{cases} \tag{4.3}$$

Algorithm 1 ensures perfect classification of independent test data under the HDLSS asymptotic regime by Theorem 8. In Algorithm 1, we also estimate a bias term as $g_{j,\mathrm{SMDP}}$ for each repetition. In fact, we do not need to estimate this term since projections of $\mathcal{X}_{te}$ onto w_{SMDP} converges two distinct points for each class, one for each class. In Algorithm 2, we simply achieve a threshold for binary classification of this one-dimensional well-separated data by using Linear Discriminant Analysis (LDA) by Fisher [1936]. Taking this approach eliminates the need to estimate m_1 and m_2. A detailed algorithm is given in Algorithm 2.

Chapter 5

Simulation

In this chapter, we numerically show that $\phi_{\text{SMDP-I}}$ in (4.2) and $\phi_{\text{SMDP-II}}$ in (4.3) can achieve asymptotic perfect classification under various heterogeneous covariance models. We compare classification rates of SMDP algorithms with several other classification rules, which are the maximal data piling classification rule (MDP) by Ahn and Marron [2010], the projected ridge classification rule (PRD) by Chang et al. [2021], Distance Weighted Discrimination (DWD) by Marron et al. [2007], Transformed Distance-Based Discriminant Analysis (T-DBDA) by Aoshima and Yata [2019] and Transformed Geometrical Quadratic Discriminant Analysis (T-GQDA) by Ishii et al. [2022].

Our model, which assumed to be satisfying Assumptions 1—5, is that $X_{kj} \sim \mathcal{N}_p(\boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})$ for $k = 1, 2$, $j = 1, \ldots, 20$ and $p = 10,000$. We set $\boldsymbol{\mu}_{(1)} = p^{-1/2}(\sqrt{8}\mathbf{1}_{p/8}^\top, \mathbf{0}_{7p/8}^\top)^\top$, $\boldsymbol{\mu}_{(2)} = \mathbf{0}_p$. Note that in this case $\delta^2 = 1$. $\boldsymbol{\Sigma}_{(1)}$ and $\boldsymbol{\Sigma}_{(2)}$ will be given differently for each setting.

In Setting I, we assume two population have the common covariance ma-

trix, that is,

$$\mathbf{\Sigma}_{(1)} = \mathbf{\Sigma}_{(2)} = \sum_{i=1}^{2} \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top + \tau^2 \mathbf{I}_p$$

where $(\sigma_1^2, \sigma_2^2) = (20p, 10p)$,

$$(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{\sqrt{p}} \begin{bmatrix} \sqrt{2}\mathbf{1}_{p/4} & \mathbf{0}_{p/4} \\ \sqrt{2}\mathbf{1}_{p/4} & \mathbf{0}_{p/4} \\ \mathbf{0}_{p/4} & \sqrt{2}\mathbf{1}_{p/4} \\ \mathbf{0}_{p/4} & \sqrt{2}\mathbf{1}_{p/4} \end{bmatrix}$$

and $\tau^2 = 30$. In Setting II, we assume heterogeneous covariance models with equal tail eigenvalues, that is, $\tau_1^2 = \tau_2^2 =: \tau^2$. To be specific, we assume

$$\mathbf{\Sigma}_{(1)} = \sum_{i=1}^{3} \sigma_{(1),i}^2 \mathbf{u}_{(1),i} \mathbf{u}_{(1),i}^\top + \tau_1^2 \mathbf{I}_p \tag{5.1}$$

and

$$\mathbf{\Sigma}_{(2)} = \sum_{i=1}^{3} \sigma_{(2),i}^2 \mathbf{u}_{(2),i} \mathbf{u}_{(2),i}^\top + \tau_2^2 \mathbf{I}_p \tag{5.2}$$

where $(\sigma_{(1),1}^2, \sigma_{(1),2}^2, \sigma_{(1),3}^2) = (\sigma_{(2),1}^2, \sigma_{(2),2}^2, \sigma_{(2),3}^2) = (20p, 10p, 5p)$,

$$(\mathbf{u}_{(1),1}, \mathbf{u}_{(1),2}, \mathbf{u}_{(1),3}) = \frac{1}{\sqrt{p}} \begin{bmatrix} \sqrt{2}\mathbf{1}_{p/4} & \mathbf{0}_{p/4} & \mathbf{1}_{p/4} \\ \sqrt{2}\mathbf{1}_{p/4} & \mathbf{0}_{p/4} & -\mathbf{1}_{p/4} \\ \mathbf{0}_{p/4} & \sqrt{2}\mathbf{1}_{p/4} & \mathbf{1}_{p/4} \\ \mathbf{0}_{p/4} & \sqrt{2}\mathbf{1}_{p/4} & -\mathbf{1}_{p/4} \end{bmatrix},$$

$$(\mathbf{u}_{(2),1}, \mathbf{u}_{(2),2}, \mathbf{u}_{(2),3}) = \frac{1}{\sqrt{p}} \begin{bmatrix} \mathbf{1}_{p/4} & \sqrt{2}\mathbf{1}_{p/4} & \mathbf{0}_{p/4} \\ \mathbf{1}_{p/4} & \mathbf{0}_{p/4} & \sqrt{2}\mathbf{1}_{p/4} \\ \mathbf{1}_{p/4} & -\sqrt{2}\mathbf{1}_{p/4} & \mathbf{0}_{p/4} \\ \mathbf{1}_{p/4} & \mathbf{0}_{p/4} & -\sqrt{2}\mathbf{1}_{p/4} \end{bmatrix}$$

and $\tau_1^2 = \tau_2^2 =: \tau^2 = 30$. In Setting III, we assume heterogeneous covariance models with unequal tail eigenvalues, that is, $\tau_1^2 > \tau_2^2$. We continue to assume $\mathbf{\Sigma}_{(1)}$ in (5.1) and $\mathbf{\Sigma}_{(2)}$ in (5.2), but $\tau_1^2 = 30$ and $\tau_2^2 = 15$ for Setting III. Note that in Settings II and III, $m > \max(m_1, m_2)$ in contrast to Setting I where $m = m_1 = m_2$.

To clearly check classification performances of each classification rule, we use the true numbers of m_1, m_2 and m for $\phi_{\text{SMDP-I}}$ and $\phi_{\text{SMDP-II}}$. Also, we use the true number of strongly spiked eigenvalues for T-DBDA and T-GQDA. For $\phi_{\text{SMDP-I}}$ and $\phi_{\text{SMDP-II}}$, we set $n_{1,te} = n_{2,te} = 6$ so that $\mathcal{X}_{te}$ consists of 30% of original training data $\mathcal{X}$. Also, we set $K = 10$ in Algorithms 1 and 2. The classification rates are obtained using $1,000$ independent observations (500 independent observations for each class). We repeat this procedure 100 times and average classification rates to estimate classification accuracy of each classification rule.

Table 5.1 shows all simulation results from Setting I to Setting III. We remark that PRD by Chang et al. [2021] yields nearly perfect classification not only in case of $\mathbf{\Sigma}_{(1)} = \mathbf{\Sigma}_{(2)}$ but also in case of $\mathbf{\Sigma}_{(1)} \neq \mathbf{\Sigma}_{(2)}$ and $\tau_1^2 = \tau_2^2$. However, this classification rule achieves poor classification performances when $\tau_1^2 \neq \tau_2^2$. In contrast, we can check that $\phi_{\text{SMDP-I}}$ and $\phi_{\text{SMDP-II}}$ achieve nearly perfect classification in all of the settings. These results confirm that our approach, projecting w_{MDP} onto the nullspace of the common leading eigenspace, successfully works under various heterogeneous covariance models.

Setting	$\phi_{\text{SMDP-I}}$	$\phi_{\text{SMDP-II}}$	MDP	PRD	DWD	T-DBDA	T-GQDA
I	0.999	0.999	0.859	0.999	0.701	0.784	0.678
	(0.001)	(0.001)	(0.102)	(0.001)	(0.097)	(0.101)	(0.083)
II	0.975	0.974	0.728	0.982	0.622	0.654	0.870
	(0.014)	(0.014)	(0.076)	(0.008)	(0.055)	(0.055)	(0.042)
III	0.993	0.993	0.669	0.576	0.620	0.654	0.991
	(0.005)	(0.006)	(0.004)	(0.004)	(0.056)	(0.060)	(0.012)

Table 5.1 Estimates of the classification accuracy of Setting I to Setting III are given in the first row of each cell, and standard errors are given in the second row of each cell.

Chapter 6

Discussion

In this work, we proposed Second Maximal Data Piling (SMDP) algorithms, which estimate a second maximal data piling direction by projecting w_{MDP} onto the nullspace of the common leading eigenspace, based on a data-splitting approach, and compute discrimination rules based on the estimated directions. The resulting classifiers can achieve asymptotic perfect classification for generalized heterogeneous spiked covariance models.

There has been relatively scarce works on a binary classification problem for cases where $\mathbf{\Sigma}_{(1)}$ or $\mathbf{\Sigma}_{(2)}$ has strong spikes, which reflects much more realistic and interesting situations for HDLSS data. Aoshima and Yata [2019] proposed a distance-based classifier, while Ishii et al. [2022] proposed geometrical quadratic discriminant analysis for this problem. Both assumed not only the dimension of data p but also training sample sizes of each class n_1 and n_2 tend to infinity to achieve perfect classification. Ishii [2020] proposed another distance-based classifier which achieves perfect classification even when n_1 and n_2 are fixed, but limited to the one-component covariance model (with

$m_1 = m_2 = 1$). All of these works were based on a data transformation technique, which essentially projecting the independent test data onto the nullspace of the leading eigenspace. Our results were also based on a similar idea of removing the leading eigenspace, but we further suggested the concept of double data piling phenomenon and revealed the relationship between the maximal data piling direction of training data and the second maximal data piling direction of independent test data under generalized heterogeneous spiked covariance models.

Bibliography

J. Ahn and J. S. Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, 2010.

J. Ahn, J. S. Marron, K. M. Muller, and Y.-Y. Chi. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766, 2007.

M. Aoshima and K. Yata. Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics*, 71(3):473–503, 2019.

R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

W. Chang, J. Ahn, and S. Jung. Double data piling leads to perfect classification. *Electronic Journal of Statistics*, 15(2):6382–6428, 2021.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

A. Ishii. A classifier under the strongly spiked eigenvalue model in high-dimension, low-sample-size context. *Communications in Statistics - Theory and Methods*, 49(7):1561–1577, 2020.

A. Ishii, K. Yata, and M. Aoshima. Geometric classifiers for high-dimensional noisy data. *Journal of Multivariate Analysis*, 188:104850, 2022.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

S. Jung. Continuum directions for supervised dimension reduction. *Computational Statistics & Data Analysis*, 125:27–43, 2018.

S. Jung and J. S. Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

S. Jung, A. Sen, and J. Marron. Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis*, 109: 190–203, 2012.

S. Jung, M. H. Lee, and J. Ahn. On the number of principal components in high dimensions. *Biometrika*, 105(2):389–402, 2018.

A. N. Kolmogorov and Y. A. Rozanov. On strong mixing conditions for stationary gaussian processes. *Theory of Probability & Its Applications*, 5(2): 204–208, 1960.

S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.

M. H. Lee, J. Ahn, and Y. Jeon. HDLSS discrimination with adaptive data

piling. *Journal of Computational and Graphical Statistics*, 22(2):433–451, 2013.

J. T. Leek. Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics*, 67(2):344–352, 2010.

J. S. Marron, M. J. Todd, and J. Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

D. Passemier and J. Yao. Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, 127:173–183, 2014.

X. Qiao, H. H. Zhang, Y. Liu, M. J. Todd, and J. S. Marron. Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, 105(489):401–414, 2010.

D. Shen, H. Shen, and J. S. Marron. A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, 17 (150):1–34, 2016.

Appendix A

Asymptotic Properties of High-dimensional Sample Within-scatter Matrix

For any vector $v \in \mathbb{R}^l$ ($l \in \mathbb{N}$), let $[v]_i$ denote the ith element of v. For any matrix $\mathbf{M} \in \mathbb{R}^{l \times l'}$ ($l, l' \in \mathbb{N}$), let $[\mathbf{M}]_i$ and $[\mathbf{M}]^j$ denote the ith row and the jth column of $\mathbf{M}$, respectively. Also, let $[\mathbf{M}]_{i,j}$ denote the (i,j)-coordinate of $\mathbf{M}$. Let $\mathbf{1}_l \in \mathbb{R}^l$ (and $\mathbf{0}_l \in \mathbb{R}^l$) denote a vector whose all entries are 1 (and 0, respectively). Write an $(l \times l)$ identity matrix as $\mathbf{I}_l$, and an $(l \times l')$ matrix whose entries are all zero as $\mathbf{O}_{l \times l'}$.

Recall that the matrix of true principal component scores of $\mathbf{X}_k$ is $\mathbf{Z}_{(k)} = \mathbf{\Lambda}_{(k)}^{-\frac{1}{2}} \mathbf{U}_{(k)}^\top (\mathbf{X}_k - \boldsymbol{\mu}_{(k)} \mathbf{1}_{n_k}^\top) = [z_{(k),1}, \ldots, z_{(k),p}]^\top \in \mathbb{R}^{p \times n_k}$ where $z_{(k),j}$ is a vector of jth principal component scores of the kth class. We write $\bar{z}_{(k),i} = n_k^{-1} z_{(k),i}^\top \mathbf{1}_{n_k}$. Also, denote a vector of true principal component scores of independent observation Y by $\zeta = (\zeta_1, \ldots, \zeta_p)^\top$. Note that each element of $\mathbf{Z}_{(k)}$ and ζ is uncorrelated, and has mean zero and unit variance.

The following lemma follows directly from Lemma C.1. of Chang et al.

[2021].

Lemma 10 *Suppose Assumptions 1—5 hold. For $k = 1, 2$, the following hold as $p \to \infty$.*

(i) $p^{-1}\boldsymbol{\mu}^\top \mathbf{U}_{(k)} \boldsymbol{\Lambda}_{(k)}^{1/2} \zeta \xrightarrow{P} \sum_{i=1}^{m_k} \sigma_{(k),i} \cos\theta_{(k),i} \delta\zeta_i$

(ii) $p^{-1}\boldsymbol{\mu}^\top \mathbf{U}_{(k)} \boldsymbol{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)} \xrightarrow{P} \sum_{i=1}^{m_k} \sigma_{(k),i} \cos\theta_{(k),i} \delta z_{(k),i}^\top$

(iii) $p^{-1}\mathbf{Z}_{(k)}^\top \boldsymbol{\Lambda}_{(k)} \zeta \xrightarrow{P} \sum_{i=1}^{m_k} \sigma_{(k),i}^2 z_{(k),i} \zeta_i$

(iv) $p^{-1}\mathbf{Z}_{(k)}^\top \boldsymbol{\Lambda}_{(k)} \mathbf{Z}_{(k)} \xrightarrow{P} \sum_{i=1}^{m_k} \sigma_{(k),i}^2 z_{(k),i} z_{(k),i}^\top + \tau_k^2 \mathbf{I}_{n_k}$

From now on, we examine asymptotic properties of the sample within-scatter matrix $\mathbf{S}_W = (\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^\top = \sum_{i=1}^{n-2} \hat{\lambda}_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top$. Since the dimension of $\mathbf{S}_W$ grows as $p \to \infty$, we instead use the $n \times n$ dual matrix, $\mathbf{S}_D = (\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}})$, which shares its nonzero eigenvalues with $\mathbf{S}_W$. We write the singular-value-decomposition of $\mathbf{X} - \bar{\mathbf{X}} = \hat{\mathbf{U}}_1 \mathbf{D}_1 \hat{\mathbf{V}}_1^\top = \sum_{i=1}^{n-2} d_i \hat{\mathbf{u}}_i \hat{\mathbf{v}}_i^\top$, where $\hat{\mathbf{u}}_i$ is the ith eigenvector of $\mathbf{S}_W$, d_i is the ith largest nonzero singular value, and $\hat{\mathbf{v}}_i$ is the vector of normalized sample principal component scores. Write $\hat{\mathbf{v}}_i = (\hat{\mathbf{v}}_{i,1}^\top, \hat{\mathbf{v}}_{i,2}^\top)^\top$ where $\hat{\mathbf{v}}_{i,1} \in \mathbb{R}^{n_1}$ and $\hat{\mathbf{v}}_{i,2} \in \mathbb{R}^{n_2}$. Then for $1 \le i \le n-2$, we can write

$$\hat{\mathbf{u}}_i = d_i^{-1}(\mathbf{X} - \bar{\mathbf{X}})\hat{\mathbf{v}}_i = \hat{\lambda}_i^{-1/2} \sum_{k=1}^{2} \mathbf{U}_{(k)} \boldsymbol{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)} (\mathbf{I}_{n_k} - \frac{1}{n_k} \mathbf{J}_{n_k}) \hat{\mathbf{v}}_{i,k}. \tag{A.1}$$

Recall that $\mathbf{W}_{(k)} = [\sigma_{(k),1} z_{(k),1}, \ldots, \sigma_{(k),m_k} z_{(k),m_k}]$ is a $n_k \times m_k$ matrix of the leading m_k principal component scores of the kth class for each $k = 1, 2$.

Lemma 11 *Suppose Assumptions 1—5 hold. Then,*

$$p^{-1}\mathbf{S}_D \xrightarrow{P} \mathbf{S}_0 = \begin{pmatrix} \mathbf{S}_{0,11} & \mathbf{S}_{0,12} \\ \mathbf{S}_{0,21} & \mathbf{S}_{0,22} \end{pmatrix}$$

as $p \to \infty$ *where*

$$\mathbf{S}_{0,ii} = (\mathbf{I}_{n_i} - \frac{1}{n_i}\mathbf{J}_{n_i})(\mathbf{W}_{(i)}\mathbf{W}_{(i)}^\top + \tau_i^2\mathbf{I}_{n_i})(\mathbf{I}_{n_i} - \frac{1}{n_i}\mathbf{J}_{n_i})$$

for $i = 1, 2$ *and*

$$\mathbf{S}_{0,ij} = (\mathbf{I}_{n_i} - \frac{1}{n_i}\mathbf{J}_{n_i})(\mathbf{W}_{(i)}\mathbf{R}_{(i)}^\top\mathbf{R}_{(j)}\mathbf{W}_{(j)}^\top)(\mathbf{I}_{n_j} - \frac{1}{n_j}\mathbf{J}_{n_j})$$

for $1 \le i \ne j \le 2$.

Proof Observe that $\mathbf{X} - \bar{\mathbf{X}} = \mathbf{X}(\mathbf{I}_n - \mathbf{J}) = [\mathbf{U}_{(1)}\mathbf{\Lambda}_{(1)}^{1/2}\mathbf{Z}_{(1)}\ \mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)}](\mathbf{I}_n - \mathbf{J})$. Then we can write

$$\frac{\mathbf{S}_D}{p} = (\mathbf{I}_n - \mathbf{J})\begin{pmatrix} p^{-1}\mathbf{Z}_{(1)}^\top\mathbf{\Lambda}_{(1)}\mathbf{Z}_{(1)} & p^{-1}\mathbf{Z}_{(1)}^\top\mathbf{\Lambda}_{(1)}^{1/2}\mathbf{U}_{(1)}^\top\mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)} \\ p^{-1}\mathbf{Z}_{(2)}^\top\mathbf{\Lambda}_{(2)}^{1/2}\mathbf{U}_{(2)}^\top\mathbf{U}_{(1)}\mathbf{\Lambda}_{(1)}^{1/2}\mathbf{Z}_{(1)} & p^{-1}\mathbf{Z}_{(2)}^\top\mathbf{\Lambda}_{(2)}\mathbf{Z}_{(2)} \end{pmatrix}(\mathbf{I}_n - \mathbf{J}).$$

By Lemma 10 (d), we have $p^{-1}\mathbf{Z}_{(i)}^\top\mathbf{\Lambda}_{(i)}\mathbf{Z}_{(i)} \xrightarrow{P} \mathbf{W}_{(i)}\mathbf{W}_{(i)}^\top + \tau_i^2\mathbf{I}_{n_i}$ as $p \to \infty$. Thus, it suffices to show that $p^{-1}\mathbf{Z}_{(1)}^\top\mathbf{\Lambda}_{(1)}^{1/2}\mathbf{U}_{(1)}^\top\mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)} \xrightarrow{P} \mathbf{W}_{(1)}\mathbf{R}_{(1)}^\top\mathbf{R}_{(2)}\mathbf{W}_{(2)}^\top$ as $p \to \infty$. Write $\mathbf{U}_{(k),2} = [\mathbf{u}_{(k),m_k+1}, \ldots, \mathbf{u}_{(k),p}]$ so that $\mathbf{U}_{(k)} = [\mathbf{U}_{(k),1}\ \mathbf{U}_{(k),2}]$. Also, write $\mathbf{\Lambda}_{(k),1} = \mathrm{Diag}(\lambda_{(k),1}, \ldots, \lambda_{(k),m_k})$ and $\mathbf{\Lambda}_{(k),2} = \mathrm{Diag}(\lambda_{(k),m_k+1}, \ldots, \lambda_{(k),p})$. Finally, write $\mathbf{Z}_{(k),1} = [z_{(k),1}, \ldots, z_{(k),m_k}]^\top$ and $\mathbf{Z}_{(k),2} = [z_{(k),m_k+1}, \ldots, z_{(k),p}]^\top$ so that $\mathbf{Z}_{(k)} = \begin{pmatrix} \mathbf{Z}_{(k),1} \\ \mathbf{Z}_{(k),2} \end{pmatrix}$. Then, we can decompose $\mathbf{Z}_{(1)}^\top\mathbf{\Lambda}_{(1)}^{1/2}\mathbf{U}_{(1)}^\top\mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)} = \sum_{i,j=1}^2 \mathbf{A}_{ij}$ where $\mathbf{A}_{ij} = \mathbf{Z}_{(1),i}^\top\mathbf{\Lambda}_{(1),i}^{1/2}\mathbf{U}_{(1),i}^\top\mathbf{U}_{(2),j}\mathbf{\Lambda}_{(2),j}^{1/2}\mathbf{Z}_{(2),j}$ for $i, j = 1, 2$. We claim that (a) $p^{-1}\mathbf{A}_{12}, p^{-1}\mathbf{A}_{21}, p^{-1}\mathbf{A}_{22} \xrightarrow{P} \mathbf{O}_{n_1 \times n_2}$ and (b) $p^{-1}\mathbf{A}_{11} \xrightarrow{P} \mathbf{W}_{(1)}\mathbf{R}_{(1)}^\top\mathbf{R}_{(2)}\mathbf{W}_{(2)}^\top$ as $p \to \infty$. We first prove the claim (a). By Assumption 2, there exists $M_k < \infty$ such that $\tau_{(k),i} \le M_k$ for all i. Thus,

$$\begin{aligned} p^{-2}\mathbb{E}\left\|\mathbf{A}_{12}\right\|_F^2 &= p^{-2}n_1n_2\mathrm{trace}(\mathbf{U}_{(2),2}^\top\mathbf{U}_{(1),1}\mathbf{\Lambda}_{(1),1}\mathbf{U}_{(1),1}^\top\mathbf{U}_{(2),2}\mathbf{\Lambda}_{(2),2}) \\ &= p^{-2}n_1n_2\sum_{l=1}^{m_1}\sum_{l'=m_2+1}^{p}(p\sigma_{(1),l}^2 + \tau_{(1),l}^2)\tau_{(2),l'}^2(\mathbf{u}_{(1),l}^\top\mathbf{u}_{(2),l'})^2 \end{aligned}$$

$$\leq p^{-2} n_1 n_2 m_1 (p\sigma_{(1),1}^2 + M_1^2) M_2^2 \to 0$$

as $p \to \infty$ where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Thus, $p^{-1}\mathbf{A}_{12} \xrightarrow{P} \mathbf{O}_{n_1 \times n_2}$ as $p \to \infty$, and $p^{-1}\mathbf{A}_{21} \xrightarrow{P} \mathbf{O}_{n_1 \times n_2}$ can be shown in a similar manner. Similarly, we can show that $p^{-2}\mathbb{E}\,\|\mathbf{A}_{22}\|_F^2 \to 0$ and $p^{-1}\mathbf{A}_{22} \xrightarrow{P} \mathbf{O}_{n_1 \times n_2}$ as $p \to \infty$ and we complete the proof of the claim (a). The claim (b) is easily proved by the fact that $p^{-1}\mathbf{A}_{11} = p^{-1}\mathbf{Z}_{(1),1}^\top \mathbf{\Lambda}_{(1),1}^{1/2} \mathbf{R}_{(1)}^{(p)\top} \mathbf{R}_{(2)}^{(p)} \mathbf{\Lambda}_{(2),1}^{1/2} \mathbf{Z}_{(2),1} \xrightarrow{P} \mathbf{W}_{(1)} \mathbf{R}_{(1)}^\top \mathbf{R}_{(2)} \mathbf{W}_{(2)}^\top$ as $p \to \infty$. ■

For the proof of Lemma 3 and Lemma 4, we will use the fact that $d_i \xrightarrow{P} \sqrt{\phi_i(\mathbf{S}_0)}$, $\hat{\mathbf{v}}_i \xrightarrow{P} v_i(\mathbf{S}_0)$ for $1 \leq i \leq n-2$ as $p \to \infty$. Recall that for any square matrix $\mathbf{M}$, $\phi_i(\mathbf{M})$ and $v_i(\mathbf{M})$ denote the ith largest eigenvalue of $\mathbf{M}$ and its corresponding eigenvector, respectively. Also, let $v_{ij}(\mathbf{M})$ be the jth coefficient of $v_i(\mathbf{M})$. We write $v_i(\mathbf{S}_0) = (\tilde{v}_{i1}(\mathbf{S}_0)^\top, \tilde{v}_{i2}(\mathbf{S}_0)^\top)^\top$ where $\tilde{v}_{i1}(\mathbf{S}_0) \in \mathbb{R}^{n_1}$ and $\tilde{v}_{i2}(\mathbf{S}_0) \in \mathbb{R}^{n_2}$. Also, write $v_i(\mathbf{\Phi}_{\tau_1,\tau_2}) = (\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2})^\top, \tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2})^\top)^\top$ where $\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2}) \in \mathbb{R}^{m_1}$ and $\tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2}) \in \mathbb{R}^{m_2}$.

A.1 Proof of Lemma 3

Proof Recall that $\mathbf{S}_W$ shares its nonzero eigenvalues with $\mathbf{S}_D$, and since ϕ_i is a continuous function of elements of a symmetric matrix, we have $\phi_i(p^{-1}\mathbf{S}_W) \xrightarrow{P} \phi_i(\mathbf{S}_0)$ as $p \to \infty$ for $1 \leq i \leq n-2$. First, assume $\tau_1^2 = \tau_2^2 =: \tau^2$. Then, $\mathbf{S}_0 = (\mathbf{I}_n - \mathbf{J})(\mathbf{W}\mathbf{W}^\top + \tau^2 \mathbf{I}_n)(\mathbf{I}_n - \mathbf{J})$. In a similar way to the proof of Lemma C.2. of Chang et al. [2021], we have $\phi_i(\mathbf{S}_0) = \phi_i(\mathbf{\Phi}) + \tau^2$ for $1 \leq i \leq m$ and $\phi_i(\mathbf{S}_0) = \tau^2$ for $m+1 \leq i \leq n-2$ and Lemma 3 (i) is proved.

Next, assume $\tau_1^2 > \tau_2^2$. Then,

$$\mathbf{S}_0 = (\mathbf{I}_n - \mathbf{J}) \begin{pmatrix} \mathbf{W}_{(1)}\mathbf{W}_{(1)}^\top + \tau_1^2 \mathbf{I}_{n_1} & \mathbf{W}_{(1)}\mathbf{R}_{(1)}^\top \mathbf{R}_{(2)} \mathbf{W}_{(2)}^\top \\ \mathbf{W}_{(2)}\mathbf{R}_{(2)}^\top \mathbf{R}_{(1)} \mathbf{W}_{(1)}^\top & \mathbf{W}_{(2)}\mathbf{W}_{(2)}^\top + \tau_2^2 \mathbf{I}_{n_2} \end{pmatrix} (\mathbf{I}_n - \mathbf{J}).$$

First, let $u = (u_1^\top, \mathbf{0}_{n_2}^\top)^\top \in \mathbb{R}^n$ be an unit vector satisfying $\mathbf{W}_{(1)}^\top(\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})u_1 = \mathbf{0}_{m_1}$ and $\mathbf{1}_{n_1}^\top u_1 = 0$. Then,

$$\mathbf{S}_0 u = \begin{pmatrix} (\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})(\mathbf{W}_{(1)}\mathbf{W}_{(1)}^\top + \tau_1^2\mathbf{I}_{n_1})(\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})u_1 \\ (\mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{J}_{n_2})\mathbf{W}_{(2)}\mathbf{R}_{(2)}^\top\mathbf{R}_{(1)}\mathbf{W}_{(1)}^\top(\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})u_1 \end{pmatrix} = \tau_1^2 \begin{pmatrix} u_1 \\ \mathbf{0}_{n_2} \end{pmatrix} = \tau_1^2 u. \tag{A.2}$$

It implies that $\mathbf{S}_0$ has an eigenvalue τ_1^2 of multiplicity $(n_1 - m_1 - 1)$. Likewise, we can show that $\mathbf{S}_0$ has an eigenvalue τ_2^2 of multiplicity $(n_2 - m_2 - 1)$. Lastly, let $u_i = (u_{i1}^\top, u_{i2}^\top)^\top \in \mathbb{R}^n$ $(1 \le i \le m_1 + m_2)$ be an unit vector with $u_{i1} = (\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})\mathbf{W}_{(1)}\mathbf{\Phi}_1^{-1/2}\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2})$ and $u_{i2} = (\mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{J}_{n_2})\mathbf{W}_{(2)}\mathbf{\Phi}_2^{-1/2}\tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2})$. Then,

$$\begin{aligned} \mathbf{S}_0 u_i &= \begin{pmatrix} (\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})\mathbf{W}_{(1)}\mathbf{\Phi}_1^{-1/2}((\mathbf{\Phi}_1 + \tau_1^2\mathbf{I}_{m_1})\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2}) + \mathbf{\Phi}_1^{1/2}\mathbf{R}_{(1)}^\top\mathbf{R}_{(2)}\mathbf{\Phi}_2^{1/2}\tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2})) \\ (\mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{J}_{n_2})\mathbf{W}_{(2)}\mathbf{\Phi}_2^{-1/2}(\mathbf{\Phi}_2^{1/2}\mathbf{R}_{(2)}^\top\mathbf{R}_{(1)}\mathbf{\Phi}_1^{1/2}\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2}) + (\mathbf{\Phi}_2 + \tau_2^2\mathbf{I}_{m_2})\tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2})) \end{pmatrix} \\ &= \phi_i(\mathbf{\Phi}_{\tau_1,\tau_2}) \begin{pmatrix} (\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})\mathbf{W}_{(1)}\mathbf{\Phi}_1^{-1/2}\tilde{v}_{i1}(\mathbf{\Phi}_{\tau_1,\tau_2}) \\ (\mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{J}_{n_2})\mathbf{W}_{(2)}\mathbf{\Phi}_2^{-1/2}\tilde{v}_{i2}(\mathbf{\Phi}_{\tau_1,\tau_2}) \end{pmatrix} = \phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})u_i \end{aligned} \tag{A.3}$$

for all $1 \le i \le m_1 + m_2$. Thus, $\mathbf{S}_0$ has eigenvalues $\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})$ for $1 \le i \le m_1 + m_2$. In summary, $\mathbf{S}_0$ has eigenvalues τ_1^2 of multiplicity $(n_1 - m_1 - 1)$, τ_2^2 of multiplicity $(n_2 - m_2 - 1)$ and $\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})$ for $1 \le i \le m_1 + m_2$. Note that $\mathbf{\Phi}_{\tau_1,\tau_2}$ can be decomposed as follows:

$$\begin{aligned} \mathbf{\Phi}_{\tau_1,\tau_2} &= \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_1^{1/2}\mathbf{R}_{(1)}^\top\mathbf{R}_{(2)}\mathbf{\Phi}_2^{1/2} \\ \mathbf{\Phi}_2^{1/2}\mathbf{R}_{(2)}^\top\mathbf{R}_{(1)}\mathbf{\Phi}_1^{1/2} & \mathbf{\Phi}_2 \end{pmatrix} + \begin{pmatrix} \tau_1^2\mathbf{I}_{m_1} & \mathbf{O}_{m_1\times m_2} \\ \mathbf{O}_{m_2\times m_1} & \tau_2^2\mathbf{I}_{m_2} \end{pmatrix} \\ &:= \mathbf{\Phi}_D + \mathbf{N}. \end{aligned}$$

Since

$$\mathbf{\Phi}_D = \begin{pmatrix} \mathbf{\Phi}_1^{1/2} & \mathbf{O}_{m_1\times m_2} \\ \mathbf{O}_{m_2\times m_1} & \mathbf{\Phi}_2^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{(1)}^\top \\ \mathbf{R}_{(2)}^\top \end{pmatrix} \begin{pmatrix} \mathbf{R}_{(1)} & \mathbf{R}_{(2)} \end{pmatrix} \begin{pmatrix} \mathbf{\Phi}_1^{1/2} & \mathbf{O}_{m_1\times m_2} \\ \mathbf{O}_{m_2\times m_1} & \mathbf{\Phi}_2^{1/2} \end{pmatrix},$$

$\mathbf{\Phi}_D$ is of rank m and shares its nonzero eigenvalues with $\mathbf{\Phi}$. By Weyl's inequality, we have $\phi_{m_1}(\mathbf{\Phi}_{\tau_1,\tau_2}) \geq \phi_{m_1+m_2}(\mathbf{\Phi}_D)+\phi_{m_1}(\mathbf{N}) \geq \tau_1^2$ and $\phi_{m_1+m_2}(\mathbf{\Phi}_{\tau_1,\tau_2}) \geq \phi_{m_1+m_2}(\mathbf{\Phi}_D) + \phi_{m_1+m_2}(\mathbf{N}) \geq \tau_2^2$. Hence, if we denote $\phi_{m_1+m_2+1}(\mathbf{\Phi}_{\tau_1,\tau_2}) = 0$, then there exists k_0 $(m_1 \leq k_0 \leq m_1 + m_2)$ such that $\phi_{k_0}(\mathbf{\Phi}_{\tau_1,\tau_2}) \geq \tau_1^2 \geq \phi_{k_0+1}(\mathbf{\Phi}_{\tau_1,\tau_2})$ and we have Lemma 3 (ii). ■

A.2 Proof of Lemma 4

Proof From (A.1), we can write

$$\mathbf{u}_j^\top \hat{\mathbf{u}}_i = \left(\frac{\hat{\lambda}_i}{p}\right)^{-1/2} \sum_{k=1}^{2} \frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{U}_{(k)} \mathbf{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)} (\mathbf{I}_{n_k} - \frac{1}{n_k}\mathbf{J}_{n_k}) \hat{\mathbf{v}}_{i,k}.$$

Note that $p^{-1/2}\mathbf{u}_j^\top \mathbf{U}_{(k)} \mathbf{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)}$ can be decomposed into two terms:

$$\frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{U}_{(k)} \mathbf{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)} = \sum_{i=1}^{m_k} \frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{u}_{(k),i} \lambda_{(k),i}^{1/2} z_{(k),i}^\top + \sum_{i=m_k+1}^{p} \frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{u}_{(k),i} \lambda_{(k),i}^{1/2} z_{(k),i}^\top$$

for $1 \leq j \leq m$. The first term converges as $p \to \infty$:

$$\sum_{i=1}^{m_k} \frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{u}_{(k),i} \lambda_{(k),i}^{1/2} z_{(k),i}^\top \xrightarrow{P} [\mathbf{R}_{(k)}]_j \mathbf{W}_{(k)}^\top \tag{A.4}$$

The second term converges to zero in probability since for any $\epsilon > 0$, by Chebyshev's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{i=m_k+1}^{p} \frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{u}_{(k),i} \lambda_{(k),i}^{1/2} z_{(k),i} \right\| > \epsilon \right) \leq \frac{1}{p\epsilon^2} \mathbb{E}\left(\left\| \sum_{i=m_k+1}^{p} \mathbf{u}_j^\top \mathbf{u}_{(k),i} \tau_{(k),i} z_{(k),i} \right\|^2 \right) \\
&= \frac{1}{p\epsilon^2} \mathbb{E}\left(\sum_{i=m_k+1}^{p} (\mathbf{u}_j^\top \mathbf{u}_{(k),i})^2 \tau_{(k),i}^2 z_{(k),i}^\top z_{(k),i} + \sum_{m_k+1 \leq i \neq l \leq p} (\mathbf{u}_j^\top \mathbf{u}_{(k),i})(\mathbf{u}_j^\top \mathbf{u}_{(k),l}) \tau_{(k),i} \tau_{(k),l} z_{(k),i}^\top z_{(k),l} \right) \\
&= \frac{1}{p\epsilon^2} \mathbb{E}\left(\sum_{i=m_k+1}^{p} (\mathbf{u}_j^\top \mathbf{u}_{(k),i})^2 \tau_{(k),i}^2 z_{(k),i}^\top z_{(k),i} \right) \leq \frac{n_k M_k^2}{p\epsilon^2} \to 0
\end{aligned} \tag{A.5}$$

as $p \to \infty$. By combining (A.4) and (A.5), we have

$$\frac{1}{\sqrt{p}}\mathbf{u}_j^\top \mathbf{U}_{(k)}\mathbf{\Lambda}_{(k)}^{1/2}\mathbf{Z}_{(k)} \xrightarrow{P} [\mathbf{R}_{(k)}]_j \mathbf{W}_{(k)}^\top$$

as $p \to \infty$ for $k = 1, 2$. Hence,

$$\mathbf{u}_j^\top \hat{\mathbf{u}}_i \xrightarrow{P} \phi_i(\mathbf{S}_0)^{-1/2}\mathbf{e}_j^\top \mathbf{W}^\top(\mathbf{I}_n - \mathbf{J})v_i(\mathbf{S}_0) \tag{A.6}$$

as $p \to \infty$ for $1 \le j \le m$ where $\mathbf{e}_j \in \mathbb{R}^m$ is a vector whose jth coordinate is 1 and other coordinates are zero. Hence, if $\tau_1^2 = \tau_2^2 =: \tau^2$, then

$$\hat{\mathbf{u}}_i^\top \mathbf{u}_j \xrightarrow{P} \begin{cases} C_{i,j}, & 1 \le i \le m, \\ 0, & m+1 \le i \le n-2 \end{cases} \tag{A.7}$$

where

$$C_{i,j} := \sqrt{\frac{\phi_i(\mathbf{\Phi})}{\phi_i(\mathbf{\Phi}) + \tau^2}} v_{ij}(\mathbf{\Phi})$$

for $1 \le i, j \le m$. If $\tau_1^2 > \tau_2^2$, then

$$\hat{\mathbf{u}}_i^\top \mathbf{u}_j \xrightarrow{P} \begin{cases} D_{i,j}, & 1 \le i \le k_0, \\ 0, & k_0 + 1 \le i \le k_0 + (n_1 - m_1 - 1), \\ D_{i-(n_1-m_1-1),j}, & k_0 + (n_1 - m_1) \le i \le n_1 + m_2 - 1, \\ 0, & n_1 + m_2 \le i \le n-2 \end{cases} \tag{A.8}$$

where k_0 ($m_1 \le k_0 \le m_1 + m_2$) is defined in Lemma 3 (ii) and

$$D_{i,j} := \frac{1}{\sqrt{\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})}} \sum_{k=1}^{2} [\mathbf{R}_{(k)}]_j \mathbf{\Phi}_k^{1/2} \tilde{v}_{ik}(\mathbf{\Phi}_{\tau_1,\tau_2}) \tag{A.9}$$

for $1 \le i \le m_1 + m_2$ and $1 \le j \le m$. Note that

$$\cos\left(\text{Angle}(\hat{\mathbf{u}}_i, \mathcal{U})\right) = \frac{\hat{\mathbf{u}}_i^\top P_{\mathcal{U}} \hat{\mathbf{u}}_i}{\|\hat{\mathbf{u}}_i\| \|P_{\mathcal{U}} \hat{\mathbf{u}}_i\|} = \sqrt{\sum_{j=1}^{m} (\hat{\mathbf{u}}_i^\top \mathbf{u}_j)^2}.$$

Then the limit of $\cos\left(\text{Angle}(\hat{\mathbf{u}}_i, \mathcal{U})\right)$ can be obtained from (A.7) and (A.8). ■

Note that Lemma 4 can also be used to investigate the asymptotic behavior of $\hat{\mathbf{u}}_i^\top \mathbf{d}$, where $\mathbf{d}$ is the sample mean difference vector. Lemma 12 will be used frequently in the proof of main lemmas and theorems.

Lemma 12 *Suppose Assumptions 1—5 hold. Then conditional to* $\mathbf{W}_{(1)}$ *and* $\mathbf{W}_{(2)}$*, the following hold as* $p \to \infty$*.*

(i) If $\tau_1^2 = \tau_2^2 =: \tau^2$*, then*

$$p^{-1/2}\mathbf{d}^\top \hat{\mathbf{u}}_i \xrightarrow{P} \begin{cases} \sum_{j=1}^m r_j C_{i,j}, & 1 \le i \le m, \\ 0, & m+1 \le i \le n-2 \end{cases}$$

for $1 \le j \le m$ *where* $r_j := \cos\theta_j \delta + \sum_{k=1}^{m_1} [\mathbf{R}_{(1)}]_{j,k} \sigma_{(1),k} \bar{z}_{(1),k} - \sum_{k=1}^{m_2} [\mathbf{R}_{(2)}]_{j,k} \sigma_{(2),k} \bar{z}_{(2),k}$ *and* $C_{i,j}$ *is defined in (A.7).*

(ii) If $\tau_1^2 > \tau_2^2$*, then*

$$p^{-1/2}\mathbf{d}^\top \hat{\mathbf{u}}_i \xrightarrow{P} \begin{cases} \sum_{j=1}^m r_j D_{i,j}, & 1 \le i \le k_0, \\ 0, & k_0+1 \le i \le k_0 + (n_1 - m_1 - 1), \\ \sum_{j=1}^m r_j D_{i-(n_1-m_1-1),j}, & k_0 + (n_1 - m_1) \le i \le n_1 + m_2 - 1, \\ 0, & n_1 + m_2 \le i \le n-2 \end{cases}$$

for $1 \le j \le m$ *where* k_0 *is defined in Lemma 3 (ii),* r_j *is defined in Lemma 12 (i) and* $D_{i,j}$ *is defined in (A.9).*

Proof Observe that $\mathbf{d} = \bar{X}_1 - \bar{X}_2 = \boldsymbol{\mu} + \frac{1}{n_1}\mathbf{U}_{(1)}\boldsymbol{\Lambda}_{(1)}^{1/2}\mathbf{Z}_{(1)}\mathbf{1}_{n_1} - \frac{1}{n_2}\mathbf{U}_{(2)}\boldsymbol{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)}\mathbf{1}_{n_2}$ and

$$\frac{1}{\sqrt{p}}\mathbf{d}^\top \hat{\mathbf{u}}_i = \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top \hat{\mathbf{u}}_i + \frac{1}{n_1\sqrt{p}}\mathbf{1}_{n_1}^\top \mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)}^{1/2}\mathbf{U}_{(1)}^\top \hat{\mathbf{u}}_i - \frac{1}{n_2\sqrt{p}}\mathbf{1}_{n_2}^\top \mathbf{Z}_{(2)}^\top \boldsymbol{\Lambda}_{(2)}^{1/2}\mathbf{U}_{(2)}\hat{\mathbf{u}}_i.$$

First, by (A.1), we can write

$$\frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top \hat{\mathbf{u}}_i = \left(\frac{\hat{\lambda}_i}{p}\right)^{-1/2} \left(\frac{\boldsymbol{\mu}^\top \mathbf{U}_{(1)} \boldsymbol{\Lambda}_{(1)}^{1/2} \mathbf{Z}_{(1)}}{p} \;\; \frac{\boldsymbol{\mu}^\top \mathbf{U}_{(2)} \boldsymbol{\Lambda}_{(2)}^{1/2} \mathbf{Z}_{(2)}}{p} \right) (\mathbf{I}_n - \mathbf{J}) \hat{\mathbf{v}}_i.$$

Write $\mathbf{c} = (\cos\theta_1, \ldots, \cos\theta_m)^\top \in \mathbb{R}^m$ and $\mathbf{c}_k = (\cos\theta_{(k),1}, \ldots, \cos\theta_{(k),m_k})^\top \in \mathbb{R}^{m_k}$. Then we have

$$\frac{1}{p}\boldsymbol{\mu}^\top \mathbf{U}_{(k)} \boldsymbol{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)} \xrightarrow{P} \mathbf{c}_k^\top \mathbf{W}_{(k)}^\top \delta$$

as $p \to \infty$ from Lemma 10 (ii). Thus,

$$\frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top \hat{\mathbf{u}}_i \xrightarrow{P} \phi_i(\mathbf{S}_0)^{-1/2} \delta \mathbf{c}^\top \mathbf{W}^\top (\mathbf{I}_n - \mathbf{J}) v_i(\mathbf{S}_0) \tag{A.10}$$

as $p \to \infty$. Also, by (A.1), we can write

$$\begin{aligned} &\frac{1}{n_1 p} \mathbf{1}_{n_1}^\top \mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)}^{1/2} \mathbf{U}_{(1)}^\top \hat{\mathbf{u}}_i \\ &= \frac{1}{n_1} \left(\frac{\hat{\lambda}_i}{p}\right)^{-1/2} \mathbf{1}_{n_1}^\top \left(\frac{\mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)} \mathbf{Z}_{(1)}}{p} \;\; \frac{\mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)}^{1/2} \mathbf{U}_{(1)}^\top \mathbf{U}_{(2)} \boldsymbol{\Lambda}_{(2)}^{1/2} \mathbf{Z}_{(2)}}{p} \right) (\mathbf{I}_n - \mathbf{J}) \hat{\mathbf{v}}_i. \end{aligned}$$

From Lemma 10 (iv) and Theorem 11, we have

$$\frac{\mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)} \mathbf{Z}_{(1)}}{p} \xrightarrow{P} \mathbf{W}_{(1)} \mathbf{W}_{(1)}^\top + \tau_1^2 \mathbf{I}_{n_1} \tag{A.11}$$

and

$$\frac{\mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)}^{1/2} \mathbf{U}_{(1)}^\top \mathbf{U}_{(2)} \boldsymbol{\Lambda}_{(2)}^{1/2} \mathbf{Z}_{(2)}}{p} \xrightarrow{P} \mathbf{W}_{(1)} \mathbf{R}_{(1)}^\top \mathbf{R}_{(2)} \mathbf{W}_{(2)}^\top \tag{A.12}$$

as $p \to \infty$ for each $k = 1, 2$. Combining (A.11) and (A.12) gives

$$\begin{aligned} &\frac{1}{n_1 p} \mathbf{1}_{n_1}^\top \mathbf{Z}_{(1)}^\top \boldsymbol{\Lambda}_{(1)}^{1/2} \mathbf{U}_{(1)}^\top \hat{\mathbf{u}}_i \\ &\xrightarrow{P} \phi_i(\mathbf{S}_0)^{-1/2} n_1^{-1} \mathbf{1}_{n_1}^\top \mathbf{W}_{(1)} \mathbf{R}_{(1)}^\top \mathbf{W}^\top (\mathbf{I}_n - \mathbf{J}) v_i(\mathbf{S}_0) \end{aligned} \tag{A.13}$$

as $p \to \infty$. Similarly, we have

$$\begin{aligned} &\frac{1}{n_2 p} \mathbf{1}_{n_2}^\top \mathbf{Z}_{(2)}^\top \boldsymbol{\Lambda}_{(2)}^{1/2} \mathbf{U}_{(2)}^\top \hat{\mathbf{u}}_i \\ &\xrightarrow{P} \phi_i(\mathbf{S}_0)^{-1/2} n_2^{-1} \mathbf{1}_{n_2}^\top \mathbf{W}_{(2)} \mathbf{R}_{(2)}^\top \mathbf{W}^\top (\mathbf{I}_n - \mathbf{J}) v_i(\mathbf{S}_0) \end{aligned} \tag{A.14}$$

as $p \to \infty$. Hence, by combining (A.10), (A.13) and (A.14), we have

$$\begin{aligned}
&\frac{1}{\sqrt{p}}\mathbf{d}^\top \hat{\mathbf{u}}_i \\
&\xrightarrow{P} \phi_i(\mathbf{S}_0)^{-1/2}(\delta\mathbf{c}^\top + n_1^{-1}\mathbf{1}_{n_1}^\top \mathbf{W}_{(1)}\mathbf{R}_{(1)}^\top - n_2^{-1}\mathbf{1}_{n_2}^\top \mathbf{W}_{(2)}\mathbf{R}_{(2)}^\top)\mathbf{W}^\top(\mathbf{I}_n - \mathbf{J})v_i(\mathbf{S}_0)
\end{aligned}$$

as $p \to \infty$ and this completes the proof. ■

Appendix B

Technical Details of Main Results

In this section, we give the proofs of main theorems. Unless otherwise stated, we only give the proofs for the case of $\tau_1^2 > \tau_2^2$ and $m > m_1$. The proofs for the other cases are quite similar to, but much simpler than, those for this case.

B.1 Proof of Theorem 5

Proof For $Y \in \mathcal{Y}$, assume $\pi(Y) = 1$. Recall that in this case,

$$\mathcal{D} = \{1, \ldots, m_1 + m_2, n_1, \ldots, n_1 + m_2 - 1\}$$

and $\mathcal{S} = \text{span}(\{\hat{\mathbf{u}}_i\}_{i \in \mathcal{D}}, w_{\text{MDP}})$. Also, for given training dataset $\mathcal{X}$, let

$$\mathcal{D}' = \{i : 1 \leq i \leq k_0, k_0 + (n_1 - m_1) \leq i \leq n_1 + m_2 - 1\} \tag{B.1}$$

where k_0 is defined in Lemma 3 (ii). That is, $\cos(\text{Angle}(\hat{\mathbf{u}}_i, \mathcal{U})) \xrightarrow{P} D_i > 0$ for $i \in \mathcal{D}'$ and $\cos(\text{Angle}(\hat{\mathbf{u}}_i, \mathcal{U})) \xrightarrow{P} 0$ for $i \notin \mathcal{D}$ as $p \to \infty$. For notational simplicity, we write $\mathcal{D}' = \{i_1, \ldots, i_{m_1+m_2}\}$ so that $i_l < i_{l'}$ if $l < l'$. Let $\mathbf{t}^0 = (t_1, \ldots, t_m)^\top \in \mathbb{R}^m$ with $t_j = \eta_2 \cos\theta_j \delta + \sum_{k=1}^{m_1} [\mathbf{R}_{(1)}]_{jk} \sigma_{(1),k} (\zeta_k - \eta_1 \bar{z}_{(1),k}) -$

$\eta_2 \sum_{k=1}^{m_2} [\mathbf{R}_{(2)}]_{jk} \sigma_{(2),k} \bar{z}_{(2),k}$ for $1 \le j \le m$ and $\boldsymbol{\nu}^0 = \mathbf{U}_{1,\mathcal{S}} \mathbf{t}^0 + \nu_1 w_{\text{MDP}} + \bar{X}_{\mathcal{S}}$. Note that $\boldsymbol{\nu}^0 \in L_1$. We claim that $\|Y_{\mathcal{S}} - \boldsymbol{\nu}^0\| \xrightarrow{P} 0$ as $p \to \infty$. For this, we need to show that (a) $\hat{\mathbf{u}}_i^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) \xrightarrow{P} 0$ for $i \in \mathcal{D}$ and (b) $w_{\text{MDP}}^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) \xrightarrow{P} 0$ as $p \to \infty$.

First, we show that (a) $\hat{\mathbf{u}}_i^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) = p^{-1/2} \hat{\mathbf{u}}_i^\top (Y - \bar{X}) - \hat{\mathbf{u}}_i^\top \mathbf{U}_{1,\mathcal{S}} \mathbf{t}^0 \xrightarrow{P} 0$ for $1 \le i \le m$ as $p \to \infty$. Note that

$$\begin{aligned}
\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_i^\top (Y - \bar{X}) &= \frac{1}{\sqrt{p}} \hat{\mathbf{u}}_i^\top (\eta_2 \boldsymbol{\mu} + \mathbf{U}_{(1)} \boldsymbol{\Lambda}_{(1)}^{1/2} (\zeta - \frac{1}{n} \mathbf{Z}_{(1)} \mathbf{1}_{n_1}) - \frac{1}{n} \mathbf{U}_{(2)} \boldsymbol{\Lambda}_{(2)}^{1/2} \mathbf{Z}_{(2)} \mathbf{1}_{n_2}) \\
&= \frac{\eta_2}{\sqrt{p}} \hat{\mathbf{u}}_i^\top \boldsymbol{\mu} + \sum_{k=1}^{m_1} \hat{\mathbf{u}}_i^\top \mathbf{u}_{(1),k} \sigma_{(1),k} (\zeta_k - \eta_1 \bar{z}_{(1),k}) - \eta_2 \sum_{k=1}^{m_2} \hat{\mathbf{u}}_i^\top \mathbf{u}_{(2),k} \sigma_{(2),k} \bar{z}_{(2),k} + o_p(1).
\end{aligned} \tag{B.2}$$

From (A.10) and Lemma 4, we have

$$\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_{i_l}^\top (Y - \bar{X}) \xrightarrow{P} \frac{1}{\sqrt{\phi_l(\boldsymbol{\Phi}_{\tau_1,\tau_2})}} \sum_{j=1}^{m} t_j \boldsymbol{\Phi}_{lj} \tag{B.3}$$

as $p \to \infty$ where $\boldsymbol{\Phi}_{lj} = \sum_{k=1}^{2} [\mathbf{R}_{(k)}]_j \boldsymbol{\Phi}_k^{1/2} \tilde{v}_{lk}(\boldsymbol{\Phi}_{\tau_1,\tau_2})$ for $1 \le l \le m_1 + m_2$ and $1 \le j \le m$. Also, from Lemma 4, we have

$$\hat{\mathbf{u}}_{i_l}^\top \mathbf{U}_{1,\mathcal{S}} \mathbf{t}^0 \xrightarrow{P} \frac{1}{\sqrt{\phi_l(\boldsymbol{\Phi}_{\tau_1,\tau_2})}} \sum_{j=1}^{m} t_j \boldsymbol{\Phi}_{lj}$$

as $p \to \infty$. Hence, $\hat{\mathbf{u}}_{i_l}^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) = p^{-1/2} \hat{\mathbf{u}}_{i_l}^\top (Y - \bar{X}) - \hat{\mathbf{u}}_{i_l}^\top \mathbf{U}_{1,\mathcal{S}} \mathbf{t}^0 \xrightarrow{P} 0$ as $p \to \infty$ for $1 \le l \le m_1 + m_2$. Similarly we can show that $\hat{\mathbf{u}}_i^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) \xrightarrow{P} 0$ as $p \to \infty$ for $i \in \mathcal{D} \setminus \mathcal{D}'$.

Next, we show that (b) $w_{\text{MDP}}^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) = p^{-1/2} w_{\text{MDP}}^\top (Y - \bar{X}) - w_{\text{MDP}}^\top \mathbf{U}_{1,\mathcal{S}} \mathbf{t}^0 - \nu_1 \xrightarrow{P} 0$ as $p \to \infty$. We decompose $p^{-1/2} w_{\text{MDP}}^\top (Y - \bar{X})$ into the two terms:

$$\begin{aligned}
\frac{1}{\sqrt{p}} w_{\text{MDP}}^\top (Y - \bar{X}) &= \frac{\sqrt{p}}{\|\hat{\mathbf{U}}_2 \hat{\mathbf{U}}_2^\top \mathbf{d}\|} \left(\frac{\mathbf{d}^\top (Y - \bar{X})}{p} - \frac{(\hat{\mathbf{U}}_1 \hat{\mathbf{U}}_1^\top \mathbf{d})^\top (Y - \bar{X})}{p} \right) \\
&= \frac{1}{\kappa_{\text{MDP}}} (K_1 - K_2)
\end{aligned}$$

where $K_1 = \mathbf{d}^\top(Y - \bar{X})/p$ and $K_2 = (\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^\top\mathbf{d})^\top(Y - \bar{X})/p$. By Lemma 10 and Lemma 11, we have

$$\begin{aligned} K_1 &= \frac{1}{p}\left(\boldsymbol{\mu} + \frac{1}{n_1}\mathbf{U}_{(1)}\boldsymbol{\Lambda}_{(1)}^{1/2}\mathbf{Z}_{(1)}\mathbf{1}_{n_1} - \frac{1}{n_2}\mathbf{U}_{(2)}\boldsymbol{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)}\mathbf{1}_{n_2}\right)^\top \\ &\quad \left(\eta_2\boldsymbol{\mu} + \mathbf{U}_{(1)}\boldsymbol{\Lambda}_{(1)}^{1/2}(\zeta - \frac{1}{n}\mathbf{Z}_{(1)}\mathbf{1}_{n_1}) - \frac{1}{n}\mathbf{U}_{(2)}\boldsymbol{\Lambda}_{(2)}^{1/2}\mathbf{Z}_{(2)}\mathbf{1}_{n_2}\right) \\ &\xrightarrow{P} \eta_2(1-\cos^2\varphi)\delta^2 - \frac{\tau_1^2 - \tau_2^2}{n} + \sum_{j=1}^m t_j r_j \end{aligned} \tag{B.4}$$

as $p \to \infty$ where r_j is defined in Lemma 12. Also, from Lemma 12,

$$\begin{aligned} K_2 &= \frac{(\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^\top\mathbf{d})^\top(Y - \bar{X})}{p} = \sum_{i=1}^m \left(\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_i^\top\mathbf{d}\right)\left(\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_i^\top(Y - \bar{X})\right) + o_p(1) \\ &\xrightarrow{P} \sum_{l=1}^{m_1+m_2}\sum_{j=1}^m\sum_{j'=1}^m \frac{1}{\phi_l(\boldsymbol{\Phi}_{\tau_1,\tau_2})} t_j r_{j'} \boldsymbol{\Phi}_{lj}\boldsymbol{\Phi}_{lj'}. \end{aligned} \tag{B.5}$$

as $p \to \infty$. Note that the limit of κ^2_{MDP} can be obtained from the limit of $p^{-1}\|\mathbf{d}\|^2$ and $p^{-1}\|\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^\top\mathbf{d}\|^2$. Then we have

$$\begin{aligned} \kappa^2_{\mathrm{MDP}} &= \frac{1}{p}\|\mathbf{d}\|^2 - \frac{1}{p}\|\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^\top\mathbf{d}\|^2 \\ &\xrightarrow{P} (1-\cos^2\varphi)\delta^2 + \frac{\tau_1^2}{n_1} + \frac{\tau_2^2}{n_2} + \sum_{j=1}^m\sum_{j'=1}^m r_j r_{j'}\left(\delta_{jj'} - \sum_{l=1}^{m_1+m_2}\frac{1}{\phi_l(\boldsymbol{\Phi}_{\tau_1\tau_2})}\boldsymbol{\Phi}_{lj}\boldsymbol{\Phi}_{lj'}\right) \\ &= (1-\cos^2\varphi)\delta^2 + \frac{\tau_1^2}{n_1} + \frac{\tau_2^2}{n_2} + \mathbf{r}^\top\left(\mathbf{I}_m - \left(\mathbf{R}_{(1)}\boldsymbol{\Phi}_1^{1/2}\ \mathbf{R}_{(2)}\boldsymbol{\Phi}_2^{1/2}\right)\boldsymbol{\Phi}_{\tau_1,\tau_2}^{-1}\begin{pmatrix}\boldsymbol{\Phi}_1^{1/2}\mathbf{R}_{(1)}^\top \\ \boldsymbol{\Phi}_2^{1/2}\mathbf{R}_{(2)}^\top\end{pmatrix}\right)\mathbf{r} \\ &=: \kappa^2 \end{aligned} \tag{B.6}$$

as $p \to \infty$ where $\mathbf{r} = (r_1, \ldots, r_m)^\top$. Note that $\kappa^2 \geq (1-\cos^2\varphi)\delta^2 + \tau_1^2/n_1 +$

$\tau_2^2/n_2 > 0$. Combining (B.4), (B.5) and (B.6) gives

$$\frac{1}{\sqrt{p}} w_{\mathrm{MDP}}^\top (Y - \bar{X}) = \frac{1}{\kappa_{\mathrm{MDP}}}(K_1 - K_2)$$
$$\xrightarrow{P} \frac{1}{\kappa}\left\{ \eta_2(1-\cos^2\varphi)\delta^2 - \frac{\tau_1^2 - \tau_2^2}{n} + \sum_{j=1}^{m}\sum_{j'=1}^{m} t_j r_{j'} \left(\delta_{jj'} - \sum_{l=1}^{m_1+m_2} \frac{1}{\phi_l(\mathbf{\Phi}_{\tau_1,\tau_2})} \mathbf{\Phi}_{lj}\mathbf{\Phi}_{lj'} \right) \right\} \tag{B.7}$$

as $p \to \infty$ where $\delta_{jj'}$ stands for the Kronecker delta. Similarly, we have

$$w_{\mathrm{MDP}}^\top \mathbf{U}_{1,\mathcal{S}} \mathbf{t}^0 = \sum_{j=1}^{m} t_j w_{\mathrm{MDP}}^\top \mathbf{u}_j = \frac{1}{\kappa_{\mathrm{MDP}}} \sum_{j=1}^{m} t_j \left\{ \frac{1}{\sqrt{p}} \mathbf{u}_j^\top \mathbf{d} - \mathbf{u}_j^\top \hat{\mathbf{U}}_1 \left(\frac{1}{\sqrt{p}} \hat{\mathbf{U}}_1^\top \mathbf{d} \right) \right\}$$
$$\xrightarrow{P} \frac{1}{\kappa} \sum_{j=1}^{m}\sum_{j'=1}^{m} t_j r_{j'} \left(\delta_{jj'} - \sum_{l=1}^{m_1+m_2} \frac{1}{\phi_l(\mathbf{\Phi}_{\tau_1,\tau_2})} \mathbf{\Phi}_{lj}\mathbf{\Phi}_{lj'} \right) \tag{B.8}$$

as $p \to \infty$. From (B.7) and (B.8), we have $w_{\mathrm{MDP}}^\top (Y_{\mathcal{S}} - \boldsymbol{\nu}^0) \xrightarrow{P} 0$ as $p \to \infty$. Hence, from (a) and (b), we have $\|Y_{\mathcal{S}} - \nu^0\| \xrightarrow{P} 0$ as $p \to \infty$ for $Y \in \mathcal{Y}$ with $\pi(Y) = 1$. Using similar arguments, we can show for $Y \in \mathcal{Y}$ with $\pi(Y) = 2$. ■

B.2 Proof of Theorem 6

Write an eigen-decomposition of $\mathbf{S}_W^* = (\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top$ by $\hat{\mathbf{U}}_1^* \hat{\mathbf{\Lambda}}_1^* \hat{\mathbf{U}}_1^{*\top}$, where $\hat{\mathbf{\Lambda}}_1^* = \mathrm{Diag}(\hat{\lambda}_1^*, \ldots, \hat{\lambda}_{n^*-2}^*)$ in which the eigenvalues are arranged in descending order and $\hat{\mathbf{U}}_1^* = [\hat{\mathbf{u}}_1^*, \ldots, \hat{\mathbf{u}}_{n^*-2}^*]$. Also, write the singular-value-decomposition of $\mathbf{Y} - \bar{\mathbf{Y}} = \hat{\mathbf{U}}_1^* \mathbf{D}_1^* \hat{\mathbf{V}}_1^{*\top} = \sum_{i=1}^{n^*-2} d_i^* \hat{\mathbf{u}}_i^* \hat{\mathbf{v}}_i^{*\top}$ where $\hat{\mathbf{u}}_i^*$ is the ith eigenvector of $\mathbf{S}_W^*$, d_i^* is the ith nonzero largest singular value, and $\mathbf{v}_i^*$ is the vector of normalized sample principal component scores. Denote true principal components scores matrix of $\mathbf{Y}_k = [Y_{k1}, \ldots, Y_{kn_k^*}]$ by $\mathbf{Z}_{(k)}^* = \mathbf{\Lambda}_{(k)}^{-1/2} \mathbf{U}_{(k)}^\top (\mathbf{Y}_k - \mathbb{E}\mathbf{Y}_k) = [z_{(k),1}^*, \ldots, z_{(k),p}^*]^\top \in \mathbb{R}^{p \times n_k^*}$ and similar to (A.1), we can write

$$\hat{\mathbf{u}}_i^* = \hat{\lambda}_i^{*-1/2} \sum_{k=1}^{2} \mathbf{U}_{(k)} \mathbf{\Lambda}_{(k)}^{1/2} \mathbf{Z}_{(k)}^* (\mathbf{I}_{n_k^*} - \frac{1}{n_k^*} \mathbf{J}_{n_k^*}) \hat{\mathbf{v}}_{i,k}^*.$$

We write $\mathbf{W}^*_{(k)} = [\sigma_{(k),1} z^*_{(k),1}, \ldots, \sigma_{(k),m_k} z^*_{(k),m_k}]$, $\mathbf{W}^{*\top} = [\mathbf{R}_{(1)}\mathbf{W}^{*\top}_{(1)}\ \mathbf{R}_{(2)}\mathbf{W}^{*\top}_{(2)}]$, $\mathbf{\Phi}^*_{(k)} = \mathbf{W}^{*\top}_{(k)}(\mathbf{I}_{n^*_k} - \frac{1}{n^*_k}\mathbf{J}^*_{n_k})\mathbf{W}^*_{(k)}$, $\mathbf{\Phi}^* = \mathbf{W}^{*\top}(\mathbf{I}_{n^*} - \mathbf{J}^*)\mathbf{W}^*$ and

$$\mathbf{\Phi}^*_{\tau_1,\tau_2} = \begin{pmatrix} \mathbf{\Phi}^*_1 + \tau_1^2\mathbf{I}_{m_1} & \mathbf{\Phi}^{*1/2}_1\mathbf{R}^\top_{(1)}\mathbf{R}_{(2)}\mathbf{\Phi}^{*1/2}_2 \\ \mathbf{\Phi}^{*1/2}_2\mathbf{R}^\top_{(2)}\mathbf{R}_{(1)}\mathbf{\Phi}^{*1/2}_1 & \mathbf{\Phi}^*_2 + \tau_2^2\mathbf{I}_{m_2} \end{pmatrix}$$

where $\mathbf{J}^* = \begin{pmatrix} \frac{1}{n^*_1}\mathbf{J}_{n^*_1} & \mathbf{O}_{n^*_1\times n^*_2} \\ \mathbf{O}_{n^*_2\times n^*_1} & \frac{1}{n^*_2}\mathbf{J}_{n^*_2} \end{pmatrix}$.

Recall that for given training dataset $\mathcal{X}$, we define $\mathcal{D}'$ in (B.1) and write $\mathcal{D}' = \{i_1, \ldots, i_{m_1+m_2}\}$ so that $i_l < i_{l'}$ if $l < l'$. Similarly, for given independent test dataset $\mathcal{Y}$, let $\mathcal{D}^*$ be an index set such that $j \in \mathcal{D}'^*$ if and only if the probability limit of $\cos(\text{Angle}(\hat{\mathbf{u}}^*_j, \mathcal{U}))$ does not degenerate. Note that the cardinality of $\mathcal{D}'^*$ is $(m_1 + m_2)$, and for notational simplicity, we write $\mathcal{D}'^* = \{j_1, \ldots, j_{m_1+m_2}\}$ so that $j_l < j_{l'}$ if $l < l'$.

Proof First, we obtain the probability limit of $\hat{\mathbf{u}}_i^\top\hat{\mathbf{u}}^*_j$. From Lemma 11 and Lemma 3 (ii), the inner product $\hat{\mathbf{u}}_i^\top\hat{\mathbf{u}}^*_j$ becomes

$$\begin{aligned}
\hat{\mathbf{u}}_i^\top\hat{\mathbf{u}}^*_j = &\left(\frac{\hat{\lambda}_i}{p}\right)^{-1/2}\left(\frac{1}{\sqrt{p}}\mathbf{U}_{(1)}\mathbf{\Lambda}^{1/2}_{(1)}\mathbf{Z}_{(1)}(\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{J}_{n_1})\hat{\mathbf{v}}_{i,1} + \frac{1}{\sqrt{p}}\mathbf{U}_{(2)}\mathbf{\Lambda}^{1/2}_{(2)}\mathbf{Z}_{(2)}(\mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{J}_{n_1})\hat{\mathbf{v}}_{i,2}\right)^\top \\
&\left(\frac{\hat{\lambda}^*_j}{p}\right)^{-1/2}\left(\frac{1}{\sqrt{p}}\mathbf{U}_{(1)}\mathbf{\Lambda}^{1/2}_{(1)}\mathbf{Z}^*_{(1)}(\mathbf{I}_{n^*_1} - \frac{1}{n^*_1}\mathbf{J}_{n^*_1})\hat{\mathbf{v}}^*_{j,1} + \frac{1}{\sqrt{p}}\mathbf{U}_{(2)}\mathbf{\Lambda}^{1/2}_{(2)}\mathbf{Z}^*_{(2)}(\mathbf{I}_{n^*_2} - \frac{1}{n^*_2}\mathbf{J}_{n^*_2})\hat{\mathbf{v}}^*_{j,2}\right) \\
&\xrightarrow{P} \frac{1}{\sqrt{\phi_i(\mathbf{S}_0)\phi_j(\mathbf{S}^*_0)}}(\mathbf{W}^\top(\mathbf{I}_n - \mathbf{J})v_i(\mathbf{S}_0))^\top(\mathbf{W}^{*\top}(\mathbf{I}_{n^*} - \mathbf{J}^*)v_j(\mathbf{S}^*_0))
\end{aligned}$$

as $p \to \infty$ where $\mathbf{S}^*_0$ is the probability limit of $p^{-1}(\mathbf{Y} - \bar{\mathbf{Y}})^\top(\mathbf{Y} - \bar{\mathbf{Y}})$. Hence, from the proof of Lemma 4, $\hat{\mathbf{u}}_i^\top\hat{\mathbf{u}}^*_j \xrightarrow{P} 0$ as $p \to \infty$ if $i \notin \mathcal{D}'$ or $j \notin \mathcal{D}'^*$. Also, for $i_l \in \mathcal{D}'$ and $j_{l'} \in \mathcal{D}'^*$, we have

$$\hat{\mathbf{u}}_{i_l}^\top\hat{\mathbf{u}}^*_{j_{l'}} \xrightarrow{P} \frac{1}{\sqrt{\phi_l(\mathbf{\Phi}_{\tau_1,\tau_2})\phi_{l'}(\mathbf{\Phi}^*_{\tau_1,\tau_2})}}\sum_{k=1}^{m}\mathbf{\Phi}_{lk}\mathbf{\Phi}^*_{l'k} \tag{B.9}$$

as $p \to \infty$ where $\mathbf{\Phi}_{lk} = \sum_{i=1}^2[\mathbf{R}_{(i)}]_k\mathbf{\Phi}^{1/2}_{(i)}\tilde{v}_{li}(\mathbf{\Phi}_{\tau_1,\tau_2})$ and $\mathbf{\Phi}^*_{l'k} = \sum_{i=1}^2[\mathbf{R}_{(i)}]_k\mathbf{\Phi}^{1/2*}_{(i)}\tilde{v}_{l'i}(\mathbf{\Phi}^*_{\tau_1,\tau_2})$.

Next, to obtain the probability limit of $w_{\mathrm{MDP}}^{\top}\hat{\mathbf{u}}_j^*$, note that

$$w_{\mathrm{MDP}}^{\top}\hat{\mathbf{u}}_j^* = \hat{\mathbf{u}}_j^{*\top}\frac{\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^{\top}\mathbf{d}}{\|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2^{\top}\mathbf{d}\|} = \frac{1}{\kappa_{\mathrm{MDP}}}\left(\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_j^{*\top}\mathbf{d} - \frac{1}{\sqrt{p}}\hat{\mathbf{u}}_j^{*\top}\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^{\top}\mathbf{d}\right).$$

Using similar arguments to the proof of Lemma 12, for $j_{l'} \in \mathcal{D}'^*$, we can show that

$$\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_{j_{l'}}^{*\top}\mathbf{d} \xrightarrow{P} \frac{1}{\sqrt{\phi_{l'}(\mathbf{\Phi}^*_{\tau_1,\tau_2})}}\sum_{k=1}^{m} r_k \mathbf{\Phi}^*_{l'k} \tag{B.10}$$

and

$$\begin{aligned}\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_{j_{l'}}^{*\top}\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^{\top}\mathbf{d} &= \sum_{i\in\mathcal{D}'}(\hat{\mathbf{u}}_{j_{l'}}^{*\top}\hat{\mathbf{u}}_i)\left(\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_i^{\top}\mathbf{d}\right) + o_p(1) \\ &\xrightarrow{P} \frac{1}{\sqrt{\phi_{l'}(\mathbf{\Phi}^*_{\tau_1,\tau_2})}}\sum_{i=1}^{m_1+m_2}\sum_{k=1}^{m}\sum_{k'=1}^{m}\frac{1}{\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})} r_{k'}\mathbf{\Phi}_{ik}\mathbf{\Phi}_{ik'}\mathbf{\Phi}^*_{l'k}\end{aligned} \tag{B.11}$$

as $p \to \infty$ where r_k is defined in Lemma 12. Combining (B.10) and (B.11) gives

$$w_{\mathrm{MDP}}^{\top}\hat{\mathbf{u}}_{j_{l'}}^* \xrightarrow{P} \frac{1}{\kappa\sqrt{\phi_{l'}(\mathbf{\Phi}^*_{\tau_1,\tau_2})}}\sum_{k=1}^{m}\sum_{k'=1}^{m}\left(\delta_{kk'} - \sum_{i=1}^{m_1+m_2}\frac{1}{\phi_i(\mathbf{\Phi}_{\tau_1,\tau_2})}\mathbf{\Phi}_{ik}\mathbf{\Phi}_{ik'}\right) r_{k'}\mathbf{\Phi}^*_{l'k} \tag{B.12}$$

as $p \to \infty$. In contrast, for $j \notin \mathcal{D}'^*$, $p^{-1/2}\hat{\mathbf{u}}_j^{*\top}\mathbf{d} \xrightarrow{P} 0$, $p^{-1/2}\hat{\mathbf{u}}_j^{*\top}\hat{\mathbf{U}}_1\hat{\mathbf{U}}_1^{\top}\mathbf{d} \xrightarrow{P} 0$ and thus $w_{\mathrm{MDP}}^{\top}\hat{\mathbf{u}}_j^* \xrightarrow{P} 0$ as $p \to \infty$.

Let $\xi_{i,j}$ be the probability limit of $\hat{\mathbf{u}}_i^{\top}\hat{\mathbf{u}}_j^*$ and $\xi_{\mathrm{MDP},j}$ be the probability limit of $w_{\mathrm{MDP}}^{\top}\hat{\mathbf{u}}_j^*$, and write $\boldsymbol{\xi}_j = (\xi_{1,j}, \ldots, \xi_{n-2,j}, \xi_{\mathrm{MDP},j})^{\top}$ for $1 \le j \le m$. Also, let $\mathbf{V} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n-2}, w_{\mathrm{MDP}}]$ and denote the probability limit of the $(n-1)\times(n-1)$ matrix $p^{-1}\mathbf{V}^{\top}\mathbf{S}_W^*\mathbf{V}$ by $\mathbf{L}$. Since $\hat{\mathbf{u}}_i^{\top}\hat{\mathbf{u}}_j^* \xrightarrow{P} 0$ and $w_{\mathrm{MDP}}^{\top}\hat{\mathbf{u}}_j^* \xrightarrow{P} 0$ as $p \to \infty$ for $j \notin \mathcal{D}'^*$, we have $\boldsymbol{\xi}_j = \mathbf{0}_{n-1}$ for $j \notin \mathcal{D}'^*$ and

$$\mathbf{L} = \sum_{l'=1}^{m_1+m_2}\phi_{l'}(\mathbf{\Phi}^*_{\tau_1,\tau_2})\boldsymbol{\xi}_{j_{l'}}\boldsymbol{\xi}_{j_{l'}}^{\top}.$$

Meanwhile, we define the $(n-2) \times m$ matrix $\mathbf{\Omega}_1$ such that

$$[\mathbf{\Omega}_1]_{i_l,j} = [\tilde{\mathbf{\Omega}}_1]_{l,j} \tag{B.13}$$

for $1 \leq l \leq m_1 + m_2$ and $1 \leq j \leq m$ where

$$[\tilde{\mathbf{\Omega}}_1]_{l,j} = \frac{1}{\sqrt{\phi_l(\mathbf{\Phi}_{\tau_1,\tau_2})}} \mathbf{\Phi}_{lj} \tag{B.14}$$

and $[\mathbf{\Omega}_1]_{i,j} = 0$ for $i \notin \mathcal{D}'$ and $1 \leq j \leq m$. Also, we define the $(n-1) \times m$ matrix $\mathbf{\Omega} = [\mathbf{\Omega}_1^\top \ \boldsymbol{\omega}_1]^\top$ where

$$\boldsymbol{\omega}_1 = \frac{1}{\kappa}(\mathbf{I}_m - \tilde{\mathbf{\Omega}}_1^\top \tilde{\mathbf{\Omega}}_1)\mathbf{r}. \tag{B.15}$$

Lastly, we define the $(m_1 + m_2) \times m$ matrix $\tilde{\mathbf{\Omega}}_1^*$ such that

$$[\tilde{\mathbf{\Omega}}_1^*]_{i,j} = \frac{1}{\sqrt{\phi_i(\mathbf{\Phi}^*_{\tau_1,\tau_2})}} \mathbf{\Phi}^*_{ij} \tag{B.16}$$

for $1 \leq i \leq m_1 + m_2$ and $1 \leq j \leq m$. Then from (B.9) and (B.12),

$$\mathbf{\Xi} = [\boldsymbol{\xi}_{j_1}, \ldots, \boldsymbol{\xi}_{j_{m_1+m_2}}] = \mathbf{\Omega}\tilde{\mathbf{\Omega}}_1^{*\top}. \tag{B.17}$$

Since both of $\mathbf{\Omega}$ and $\tilde{\mathbf{\Omega}}_1^*$ are of rank m, by Sylvester's rank inequality, we have $\text{rank}(\mathbf{L}) = \text{rank}(\mathbf{\Xi}) = \text{span}(\mathbf{\Omega}) = m$. ■

B.3 Proof of Theorem 7

Proof For any given $\{w\} \in \bar{\mathcal{A}}$, there exists $\{v\}$ such that $v \in \text{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ and $\|w - v\| \xrightarrow{P} 0$ as $p \to \infty$. Thus it suffices to show that $v^\top \mathbf{u}_j \xrightarrow{P} 0$ as $p \to \infty$. Let $v = \mathbf{V}\mathbf{b}$ such that $\mathbf{b} \in \text{span}(\hat{\mathbf{Q}}_2)$. Note that for all $1 \leq i \leq m$, $\hat{\mathbf{q}}_i$ converges to $v_i(\mathbf{L})$ in the m-dimensional subspace $\text{span}(\mathbf{\Xi}) = \text{span}(\mathbf{\Omega})$. Hence,

for $\mathbf{b} \in \text{span}(\hat{\mathbf{Q}}_2)$, $\mathbf{b}$ is asymptotically orthogonal to $\text{span}(\mathbf{\Omega})$. From Lemma 4 and (B.8), we have

$$v^\top \mathbf{u}_j = \mathbf{b}^\top \mathbf{V}^\top \mathbf{u}_j = \mathbf{b}^\top [\mathbf{\Omega}]^j + o_p(1) \xrightarrow{P} 0 \tag{B.18}$$

as $p \to \infty$. ■

B.4 Proof of Theorem 8

Proof For any given $\{w\} \in \bar{\mathcal{A}}$, we assume $w = \mathbf{V}\mathbf{a}$ where $\mathbf{a} = (a_1, \ldots, a_{n-2}, a_{\text{MDP}})^\top$ satisfies $a_{\text{MDP}} \xrightarrow{P} \psi_{\text{MDP}}$ as $p \to \infty$. Recall that there exists $\{v\}$ such that $v \in \text{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ and $\|w - v\| \xrightarrow{P} 0$ as $p \to \infty$. Write $v = \mathbf{V}\mathbf{b}$ with $\mathbf{b} = (b_1, \ldots, b_{n-2}, b_{\text{MDP}})^\top$. Then $\|w - v\| = \|\mathbf{a} - \mathbf{b}\| \xrightarrow{P} 0$ and $b_{\text{MDP}} \xrightarrow{P} \psi_{\text{MDP}}$ as $p \to \infty$. Thus, by (B.3) and (B.7),

$$\frac{1}{\sqrt{p}} w^\top (Y - \bar{X}) = \frac{1}{\sqrt{p}} v^\top (Y - \bar{X}) + o_p(1)$$

and it suffices to obtain the probability limit of $p^{-1/2} v^\top (Y - \bar{X})$. For any observation Y, which is independent to both of $\mathcal{X}$ and $\mathcal{Y}$, assume that $\pi(Y) = 1$. Combining (B.3), (B.7) and (B.18) gives

$$\begin{aligned}
\frac{1}{\sqrt{p}} v^\top (Y - \bar{X}) &= \frac{1}{\sqrt{p}} \mathbf{b}^\top \mathbf{V}^\top (Y - \bar{X}) \\
&= \sum_{j=1}^{m} t_j \mathbf{b}^\top [\mathbf{\Omega}]^j + \frac{\psi_{\text{MDP}}}{\kappa} \left(\eta_2 (1 - \cos^2 \varphi) \delta^2 - \frac{\tau_1^2 - \tau_2^2}{n} \right) + o_p(1) \\
&\xrightarrow{P} \frac{\psi_{\text{MDP}}}{\kappa} \left(\eta_2 (1 - \cos^2 \varphi) \delta^2 - \frac{\tau_1^2 - \tau_2^2}{n} \right)
\end{aligned}$$

as $p \to \infty$ where t_j is defined in the proof of Theorem 5. Similarly, we can show that

$$\frac{1}{\sqrt{p}} w^\top (Y - \bar{X}) \xrightarrow{P} \frac{\psi_{\text{MDP}}}{\kappa} \left(-\eta_1 (1 - \cos^2 \varphi) \delta^2 - \frac{\tau_1^2 - \tau_2^2}{n} \right)$$

as $p \to \infty$ for any observation Y, independent to both of $\mathcal{X}$ and $\mathcal{Y}$, with $\pi(Y) = 2$.

■

B.5 Proof of Theorem 9

Proof From Theorem 8, for $\{w\} \in \bar{\mathcal{A}}$ with $w = \mathbf{Va}$ and $a_{\text{MDP}} \xrightarrow{P} \psi_{\text{MDP}}$ as $p \to \infty$, we can check that an asymptotic distance between the two piles of independent test data is $D(w) = \kappa^{-1}\psi_{\text{MDP}}(1 - \cos^2\varphi)\delta^2$. Let $w_{\text{SMDP}} = \mathbf{Va}_{\text{SMDP}} = \|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\text{MDP}}\|^{-1}\mathbf{V}\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\text{MDP}}$ where $\mathbf{e}_{\text{MDP}} = (\mathbf{0}_{n-2}^\top, 1)^\top$. Note that

$$\mathbf{e}_{\text{MDP}}^\top \frac{\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\text{MDP}}}{\|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\text{MDP}}\|} = \|\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\text{MDP}}\|. \tag{B.19}$$

To derive the probability limit of $\|\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\text{MDP}}\|$, we characterize an orthonormal basis of $\text{span}(\mathbf{\Omega})^\perp$, which is the orthogonal complement of the $(n-m-1)$-dimensional subspace of $\text{span}(\mathbf{\Omega})$. Note that $\text{span}(\mathbf{\Omega}_1)^\perp$, which is the orthogonal complement of $\text{span}(\mathbf{\Omega}_1)$, is $(n-m-2)$-dimensional subspace and let $\left\{\boldsymbol{\psi}_{1,1}, \ldots, \boldsymbol{\psi}_{n-m-2,1}\right\}$ be an orthonormal basis of $\text{span}(\mathbf{\Omega}_1)$. Also, let $\boldsymbol{\psi}_i = (\boldsymbol{\psi}_{i,1}^\top, 0)^\top \in \mathbb{R}^{n-1}$ for all $1 \le i \le n-m-2$. Then $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{n-m-2}$ are orthogonal to each other and $\boldsymbol{\psi}_i \in \text{span}(\mathbf{\Omega})^\perp$ for all $1 \le i \le n-m-2$.

Now, let $\boldsymbol{\psi}_0 = (\boldsymbol{\psi}_{0,1}^\top, \psi_{0,\text{MDP}})^\top \in \mathbb{R}^{n-1}$ such that

$$[\boldsymbol{\psi}_{0,1}]_{i_l} = \left[\frac{\psi_{0,\text{MDP}}}{\kappa}\tilde{\mathbf{\Omega}}_1(\mathbf{I}_m - (\tilde{\mathbf{\Omega}}_1^\top\tilde{\mathbf{\Omega}}_1)^{-1})\mathbf{r}\right]_l$$

and $[\boldsymbol{\psi}_{0,1}]_i = 0$ for $i \notin \mathcal{D}'$ and

$$\psi_{0,\text{MDP}} = \frac{\kappa}{\sqrt{\kappa^2 + \mathbf{r}^\top(\mathbf{I}_m - (\tilde{\mathbf{\Omega}}_1^\top\tilde{\mathbf{\Omega}}_1)^{-1})\tilde{\mathbf{\Omega}}_1^\top\tilde{\mathbf{\Omega}}_1(\mathbf{I}_m - (\tilde{\mathbf{\Omega}}_1^\top\tilde{\mathbf{\Omega}}_1)^{-1})\mathbf{r}}}. \tag{B.20}$$

Note that $\boldsymbol{\psi}_0$ is a unit vector. Then since

$$\begin{aligned}\boldsymbol{\Omega}^\top \boldsymbol{\psi}_0 &= \boldsymbol{\Omega}_1^\top \boldsymbol{\psi}_{0,1} + \psi_{0,\mathrm{MDP}} \boldsymbol{\omega}_1 \\ &= \frac{\psi_{0,\mathrm{MDP}}}{\kappa} \tilde{\boldsymbol{\Omega}}_1^\top \tilde{\boldsymbol{\Omega}}_1 (\mathbf{I}_m - (\tilde{\boldsymbol{\Omega}}_1^\top \tilde{\boldsymbol{\Omega}}_1)^{-1})\mathbf{r} + \frac{\psi_{0,\mathrm{MDP}}}{\kappa}(\mathbf{I}_m - \tilde{\boldsymbol{\Omega}}_1^\top \tilde{\boldsymbol{\Omega}}_1)\mathbf{r} = \mathbf{0}_m,\end{aligned}$$

we have $\boldsymbol{\psi}_0 \in \mathrm{span}(\boldsymbol{\Omega})^\perp$. It is obvious that $\boldsymbol{\psi}_0$ is orthogonal to $\boldsymbol{\psi}_i$ for all $1 \le i \le n-m-2$, and thus $\{\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{n-m-2}\}$ is an orthonormal basis of $\mathrm{span}(\boldsymbol{\Omega})^\perp$. Hence,

$$\|\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\mathrm{MDP}}\| \xrightarrow{P} \sqrt{[\boldsymbol{\psi}_0]_{n-1}^2 + \sum_{i=1}^{n-m-2} [\boldsymbol{\psi}_i]_{n-1}^2} = \psi_{0,\mathrm{MDP}}$$

as $p \to \infty$ and

$$D(w_{\mathrm{SMDP}}) = \frac{\psi_{0,\mathrm{MDP}}}{\kappa}(1 - \cos^2\varphi)\delta^2 =: \upsilon(1-\cos^2\varphi)\delta^2 > 0$$

where

$$\upsilon = \left(\kappa^2 + \mathbf{r}^\top (\mathbf{I}_m - \tilde{\boldsymbol{\Omega}}_1^\top \tilde{\boldsymbol{\Omega}}_1^{-1})\tilde{\boldsymbol{\Omega}}_1^\top \tilde{\boldsymbol{\Omega}}_1 (\mathbf{I}_m - (\tilde{\boldsymbol{\Omega}}_1^\top \tilde{\boldsymbol{\Omega}}_1)^{-1})\mathbf{r}\right)^{-1/2} \tag{B.21}$$

with probability 1. For each p, let $\{w_{\mathrm{SMDP}}, \mathbf{f}_1, \ldots, \mathbf{f}_{n-m-2}\}$ be an orthonormal basis of $\mathrm{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ and $\{w_{\mathrm{SMDP}}, \mathbf{f}_1, \ldots, \mathbf{f}_{n-m-2}, g_1, \ldots, g_m\}$ be an orthonormal basis of $\mathcal{S}_X$. For $1 \le i \le n-m-2$, write $\mathbf{f}_i = \|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{a}_i\|^{-1}\mathbf{V}\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{a}_i$. To obtain $D(\mathbf{f}_i)$, we need to derive the probability limit of

$$\mathbf{e}_{\mathrm{MDP}}^\top \frac{\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{a}_i}{\|\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{a}_i\|} = \frac{\mathbf{e}_{\mathrm{MDP}}^\top \hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{a}_i}{\|\hat{\mathbf{Q}}_2^\top \mathbf{a}_i\|}. \tag{B.22}$$

Since w_{SMDP} is orthogonal to $\mathbf{f}_i$, we have

$$w_{\mathrm{SMDP}}^\top \mathbf{f}_i = \frac{\mathbf{e}_{\mathrm{MDP}}^\top \hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_2^\top \mathbf{a}_i}{\|\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\mathrm{MDP}}\|\|\hat{\mathbf{Q}}_2^\top \mathbf{a}_i\|} = 0$$

for all p. Note that $\|\hat{\mathbf{Q}}_2^\top \mathbf{e}_{\mathrm{MDP}}\|$ converges to a strictly positive random variable, thus the probability limit of (B.22) is zero and $D(\mathbf{f}_i) = 0$ for $1 \le i \le n-m-2$.

We now show that w_{SMDP} is a second maximal data piling direction. For any given $\{w\} \in \bar{\mathcal{A}}$, write $w = a_0 w_{\text{SMDP}} + \sum_{i=1}^{n-m-2} a_i \mathbf{f}_i + \sum_{i=1}^{m} b_i g_i$. Recall that for $\{w\} \in \bar{\mathcal{A}}$, there exists $\{v\}$ such that $v \in \text{span}(\mathbf{V}\hat{\mathbf{Q}}_2)$ and $\|w - v\| \xrightarrow{P} 0$ as $p \to \infty$. Hence, $b_i = o_p(1)$ for $1 \leq i \leq m$ and since $D(\mathbf{f}_i) = 0$ for $1 \leq i \leq n - m - 2$, using similar arguments in the proof of Theorem 3.3 of Chang et al. [2021], we can show that $D(w) \leq D(w_{\text{SMDP}})$ for any $\{w\} \in \bar{\mathcal{A}}$ and the equality holds when $\|w - w_{\text{SMDP}}\| \xrightarrow{P} 0$ as $p \to \infty$. ■

국문초록

본 연구에서는 이질적인 공분산 모형을 가정하는 고차원 이항 분류 문제에 대한 두 가지 데이터 파일링 현상을 구체화한다. 데이터 파일링 현상은 훈련 데이터를 방향 벡터에 사영하였을 때 각 범주마다 정확히 두 개의 다른 값을 갖는 현상을 말한다. 첫 번째 데이터 파일링 현상은 데이터의 차원 p가 표본 크기 n보다 큰 경우 항상 발생한다. 이 연구에서는 새로운 테스트 데이터의 파일링을 의미하는 두 번째 데이터 파일링 현상이 표본 크기 n은 고정되어 있을 때 데이터의 차원 p가 증가하는 점근적 상황에서 발생할 수 있음을 보인다. 또한 테스트 데이터의 두 더미 사이의 최대 점근 거리를 만드는 두 번째 최대 데이터 파일링 방향은 첫 번째 최대 데이터 파일링 방향을 공통의 선행 고유벡터로 구성되는 공간의 직교여공간에 투영하여 얻을 수 있음을 보인다. 두 번째 데이터 파일링 현상을 바탕으로, 일반화된 이질적 스파이크 공분산 모형 하에서 고차원 저표본 데이터를 완벽하게 분류할 수 있는 새로운 선형 분류 방법을 제안한다.

주요어: 고차원 저표본, 분류, 최대 데이터 파일링, 스파이크 공분산 모형, 고차원 점근 이론

학번: 2021-29052