Journal of the Korean Statistical Society (2024) 53:704–737
https://doi.org/10.1007/s42952-024-00263-6

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Double data piling: a high-dimensional solution for asymptotically perfect multi-category classification

Taehyun Kim[1,2] · Woonyoung Chang[1,3] · Jeongyoun Ahn[4] · Sungkyu Jung[1]

Received: 22 September 2023 / Accepted: 28 February 2024 / Published online: 3 April 2024

Abstract
For high-dimensional classification, interpolation of training data manifests as the data piling phenomenon, in which linear projections of data vectors from each class collapse to a single value. Recent research has revealed an additional phenomenon known as the 'second data piling' for independent test data in binary classification, providing a theoretical understanding of asymptotically perfect classification. This paper extends these findings to multi-category classification and provides a comprehensive characterization of the double data piling phenomenon. We define the maximal data piling subspace, which maximizes the sum of pairwise distances between piles of training data in multi-category classification. Furthermore, we show that a second data piling subspace that induces data piling for independent data exists and can be consistently estimated by projecting the negatively-ridged discriminant subspace onto an estimated 'signal' subspace. By leveraging this second data piling phenomenon, we propose a bias-correction strategy for class assignments, which asymptotically achieves perfect classification. The present research sheds light on benign overfitting and enhances the understanding of perfect multi-category classification of high-dimensional discrimination with a help of high-dimensional asymptotics.

Keywords Ridge estimation · Data piling · HDLSS · Multiclass classification · Bias correction

1 Introduction

High-dimension, low-sample size (HDLSS) data analysis is becoming increasingly popular in statistical machine learning. In high-dimensional classification, it is well-known that linear classifiers are usually affected by *data piling* phenomenon (Ahn & Marron, 2010; Marron et al., 2007; Huang et al., 2013; Chang et al., 2021). The data

T. Kim and W. Chang contributed equally to this work.

Extended author information available on the last page of the article

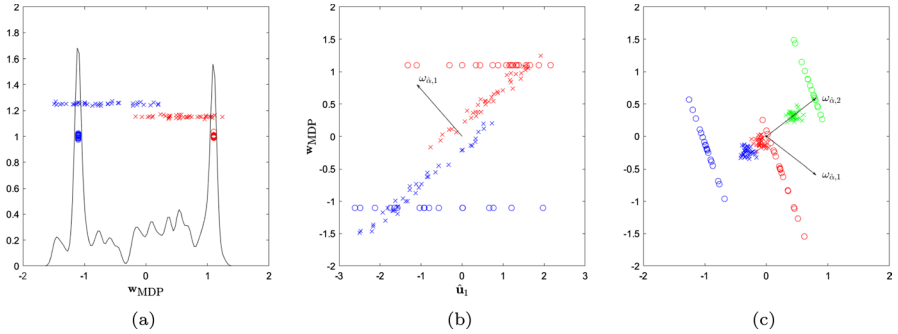


Fig. 1 Toy data example illustrating double data piling phenomenon. Circles: Training data. Crosses: Test data. Colors represent different classes. **a** The original data piling phenomenon for training data. **b** Double data piling phenomenon for both training and test data. **c** General double data piling phenomenon for multi-category classification. (Here, $\boldsymbol{\omega}_{\hat{a},i} \propto P_{\mathcal{S}}\mathbf{L}_{\hat{a},i}$; see Sect. 3.2.)

piling phenomenon occurs when projections of training data from the same class onto the normal vector of the separating hyperplane are piled on a same location in *binary* classification (Marron et al., 2007). Ahn and Marron (2010) characterized the *maximal data piling (MDP) direction*, defined as:

$$\mathbf{w}_{MDP} = \underset{\mathbf{v}\in\mathbb{R}^p : \|\mathbf{v}\|_2=1}{\operatorname{argmax}} \ \mathbf{v}'\mathbf{S}_B\mathbf{v} \quad \text{subject to} \quad \mathbf{v}'\mathbf{S}_W\mathbf{v} = 0, \tag{1}$$

where $\mathbf{S}_W$ is the $p \times p$ within-scatter matrix and $\mathbf{S}_B$ is the $p \times p$ between-scatter matrix of the training data. As long as the linear constraint in (1) is met for some $\mathbf{v} \in \mathbb{R}^d$, data piling phenomenon occurs as training data projects onto only two points on $\mathbf{v}$, one for each class. In particular, the MDP direction uniquely exists if $p > n-2$ since the rank of $\mathbf{S}_W$ is at most $\min\{p, n-2\}$. The maximal data piling direction is optimal among directions exhibiting data piling in the sense that it maximizes the distance between two piles of training data. Figure 1a shows that projections of the training data onto $\mathbf{w}_{MDP}$ are located at exactly two points, one for each class. Numerous works showed the usefulness of utilizing $\mathbf{w}_{MDP}$ in high-dimensional data analysis (Ahn & Marron, 2010; Ahn et al., 2012; Jung, 2018; Ahn et al., 2019; Chung & Ahn, 2021). On the other hand, the maximal data piling direction has been regarded as an undesirable classifier since it fits all the noise in the training data and incurs overfitting (Lee et al., 2013; Marron et al., 2007; Huang et al., 2013).

Recently, Chang et al. (2021) showed that a *second data piling* phenomenon for independent test data is possible under the HDLSS asymptotic regime where the dimension p goes to infinity while the sample size n is fixed. More noticeably, they showed that a ridged linear discriminant vector projected onto a low-dimensional subspace with a *negative* ridge parameter can yield second data piling and asymptotically achieve zero classification error. These results are in line with recent growing literature on high-dimensional linear regression models where optimal regularization can be nearly zero or even negative (Bartlett et al., 2020; Muthukumar et al., 2020; Kobak et al., 2020; Tsigler & Bartlett, 2020; Wu & Xu, 2020). Figure 1b

indicates that the projected ridged linear discriminant vector $\boldsymbol{\omega}_{\alpha,1}$ with some negative ridge parameter $\hat{\alpha} < 0$ appears to be orthogonal to the parallel lines formed by independent test data and thus yields second data piling. While Chang et al. (2021) provided compelling insights on double data piling phenomenon, their discussion was limited to the binary classification setting. In this work, we extend the findings of Chang et al. (2021) to the *multi-category* case. Our main goal is to define and estimate the subspace exhibiting second data piling for multi-category classification as shown in Fig. 1c and construct a new classification rule based on such a subspace.

1.1 Related work

Benign overfitting Recent analyses have demonstrated that *overparametrized* deep neural network models that interpolate training data can achieve successful generalization performances (Zhang et al., 2017; Belkin et al., 2019). This phenomenon, called *benign overfitting*, challenges the classical wisdom of bias-variance trade-off and has motivated theoretical and empirical research to understand why overparametrization can be beneficial (Belkin et al., 2020; Bartlett et al., 2021; Belkin, 2021). One line of work is on linear regression models (Bartlett et al., 2020; Muthukumar et al., 2020; Tsigler & Bartlett, 2020; Mei & Montanari, 2021; Hastie et al., 2022; Montanari et al., 2023; Mahdaviyeh & Naulet, 2020), showing that interpolating least squares estimator can achieve nearly perfect generalization performances. Another line of research has analyzed the asymptotic error bounds of the maximum margin classifier, demonstrating that the phenomenon of benign overfitting also occurs in binary classification (Chatterji & Long, 2021; Cao et al., 2021; Wang & Thrampoulidis, 2022). Wang et al. (2021) recently investigated sufficient conditions for multi-category classifiers under which benign overfitting occurs. We note that the majority of existing work has focused on regression or binary classification settings, which share formulaic similarities. This work will shed light on the benign overfitting phenomenon in the multi-category classification setting.

Data piling The data piling phenomenon has been a key to understanding the unique challenges in HDLSS classification. Marron et al. (2007) pointed out that the support vector machine (SVM) is likely to suffer from data piling in the HDLSS settings, resulting in poor generalization performance. Some research attempted to avoid data piling by regularizing the degrees of data piling (Lee et al., 2013; Ahn & Jeon, 2015) or using methods based on distance-weighted discrimination (DWD) (Marron et al., 2007; Huang et al., 2013; Qiao & Zhang, 2015), which was purposely developed to avoid data piling. However, recent studies have shown that, with sufficient overparametrization, hard-margin SVM can achieve good generalization even when every training point becomes a support vector, called the *support vector proliferation* (SVP) (Muthukumar et al., 2021; Hsu et al., 2022; Ardeshir et al., 2021). Moreover, they revealed that the hard-margin SVM is equivalent to the minimum-norm least squares regression under certain conditions. In this work, we provide a complete geometrical characterization of data piling as well as its extension to independent test data.

Asymptotic perfect classification It has been known that asymptotic perfect classification is possible in high or infinite dimensional classification problems if the high dimensionality is well exploited. Hall et al. (2005) showed that classical classifiers such as SVM, DWD, and the nearest neighbor classifier can achieve perfect classification under the HDLSS asymptotic regime. Delaigle and Hall (2012) showed that 'nearly' perfect classification is possible for functional data using simple linear classifiers when the standard deviation of the leading principal component scores are of the same order as, or smaller than, the population mean differences. Dai et al. (2017) proposed a functional nonparametric Bayes classifier based on the density ratios of projection scores on common eigenfunctions, and showed that it can achieve asymptotic perfect classification under certain conditions. Xue et al. (2023) suggested an optimal linear classifier for high-dimensional functional data that achieves asymptotically zero prediction error. We observe that perfect classification in the multi-category classification setting has not been addressed in the literature. Our study aims to demonstrate that asymptotic perfect classification is attainable in multi-category classification by leveraging the second data piling phenomenon.

1.2 Our contributions

We extend the discussions on the double data piling phenomenon to the general multi-category classification problem. Firstly, among subspaces exhibiting data piling, we identify the first *maximal data piling (MDP)* subspace which yields the maximum distances between piles of training data. We show that this MDP subspace is not necessarily spanned by MDP directions from binary problems. Secondly, we characterize all *second data piling (SDP)* subspaces onto which projections of independent test data from each class are piled on top of each other. We consider a common high-dimensional spiked covariance model (Johnstone, 2001; Jung & Marron, 2009; Shen et al., 2016; Wang & Fan, 2017) where several leading eigenvalues are significantly large and the rest of eigenvalues are nearly constant at $\tau^2 > 0$. We show that the SDP subspaces are asymptotically perpendicular to the leading population eigenvectors.

Among the SDP subspaces, we identify an *optimal* subspace that asymptotically induces the maximal separation between the classes. This *maximal* SDP subspace exists within $\mathcal{S}$, the 'signal' subspace that is generated by the MDP subspace and the leading sample eigenvectors. Moreover, we show that the maximal SDP subspace can be consistently estimated: A *negatively* ridged linear discriminant subspace projected on $\mathcal{S}$ yields SDP. The negative ridge phenomenon where the optimal ridge penalty can be negative was first studied by Kobak et al. (2020) using a single spiked covariance model in the overparametrized regime. We further establish that the negative ridge phenomenon also occurs in the multi-category classification setting under a general multi-spiked covariance model.

In multi-category classification, assigning class labels to the data projected onto a discriminant subspace is sometimes a non-trivial task. Pairwise comparisons or one-vs-rest comparisons are often attempted without proper justification. We present a novel classification rule for multi-category problems that corrects the asymptotic bias

in order to achieve perfect classification. The proposed bias-corrected nearest centroid classification rule considers all classes simultaneously and is theoretically shown to yield asymptotic perfect classification.

Figure 2 displays a roadmap of how we proceed to characterize the double data piling phenomenon for multi-category classification. In Sect. 2.2, we review the SDP phenomenon for binary classification. We generalize MDP from binary to multi-category classification in Sect. 2.3. Based on these foundations, we characterize the SDP subspaces for multi-category classification and propose a projected negatively-ridged discriminant subspace to estimate the maximal SDP subspace. In Sect. 4, we show that the bias-corrected classifier based on the maximal SDP subspace achieves asymptotic perfect classification. We demonstrate the theoretical findings with simulation and real data in Sect. 5. Technical details and proofs are deferred to Appendix A.

2 Preliminaries

2.1 Data models

For each $k = 1, \dots, K$, let $X_{k1}, \dots, X_{kn_k} \in \mathbb{R}^p$ be independent and identically distributed random vectors from an absolutely continuous distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_K]$ be the $p \times n$ training data matrix, where $\mathbf{X}_k = [X_{k1}, \dots, X_{kn_k}]$ is the data matrix of the kth class and $n = \sum_{k=1}^K n_k$. For class-wise sample mean vectors $\bar{X}_k = n_k^{-1} \sum_{j=1}^{n_k} X_{kj}$, let $\tilde{\mathbf{X}}_k$ be the class-wise centered data matrix of the kth class, obtained by replacing X_{kj} in $\mathbf{X}_k$ by $X_{kj} - \bar{X}_k$, and $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_K]$. Furthermore, let $\mathbf{M}$ be the $p \times K$ matrix whose kth column is $\sqrt{n_k}(\bar{X}_k - \bar{X})$ where $\bar{X} = n^{-1} \sum_{k=1}^K \sum_{j=1}^{n_k} X_{kj}$ is the total mean vector of the training data. We define the within-scatter matrix $\mathbf{S}_W = \tilde{\mathbf{X}}\tilde{\mathbf{X}}' = \sum_{k=1}^K \tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k'$ and the between-scatter matrix $\mathbf{S}_B = \mathbf{M}\mathbf{M}'$. Denote the $(n-1)$-dimensional sample space of training data by

$$\mathcal{S}_X = \text{span}(\mathbf{S}_W) + \text{span}(\mathbf{S}_B) = \text{span}\{X_{kj} - \bar{X} : k = 1, \dots, K, j = 1, \dots, n_k\}. \quad (2)$$

Assuming $p > n$, we write an eigen-decomposition of $\mathbf{S}_W$ as column space of $\mathbf{S}_W$ is of rank $n - K$ with probability one.) Also, we write the eigen-decomposition of the common covariance matrix $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$ where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_p]$ and

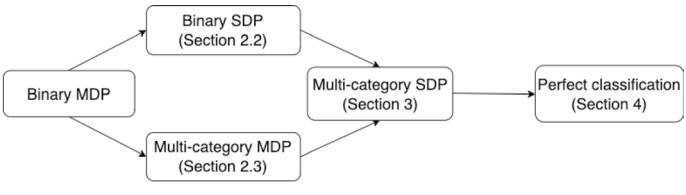


Fig. 2 Organization of the paper

$\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ arranged in descending order. We assume the following for the data model.

Assumption 1 (Non-degenerate Mean Differences) The class-wise population mean vectors $\boldsymbol{\mu}_k$ are in general position, that is, no three points are collinear. For $1 \le i < j \le K$, $\boldsymbol{\mu}_{ij} := \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and $p^{-1/2}\|\boldsymbol{\mu}_{ij}\|$ converges to a positive constant as $p \to \infty$.

Assumption 2 (Multi-spiked Covariance Model) For a fixed integer $m \ge 1$ and $\sigma_1^2 > \cdots > \sigma_m^2 > 0$, $\lambda_i = \sigma_i^2 p$ for $i = 1, \ldots, m$. The rest of the eigenvalues $\lambda_{m+1}, \ldots, \lambda_p$ are uniformly bounded, and $\sum_{i=m+1}^{p} \lambda_i / p \to \tau^2 \in (0, \infty)$.

Assumption 1 requires that the pairwise population mean differences do not degenerate as the dimension p increases (Hall et al., 2005; Qiao et al., 2010; Jung, 2018). Assumption 2 specifies the spiked covariance model for $\mathbf{\Sigma}$ where the first m eigenvalues diverge at the order of p while the rest of eigenvalues are nearly constant (Ahn et al., 2007; Jung et al., 2012; Shen et al., 2016; Jung, 2022). We call the first m eigenvalues and their corresponding eigenvectors of $\mathbf{\Sigma}$ leading eigenvalues and eigenvectors, respectively. Moreover, $\mathcal{U}_m = \text{span}(\{\mathbf{u}_i\}_{i=1}^m)$ is called the leading eigenspace.

We write the angle between two vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^p \backslash \{\mathbf{0}_p\}$ as $\text{Angle}(\mathbf{w}, \mathbf{v}) = \arccos\{\mathbf{w}'\mathbf{v}/(\|\mathbf{w}\|_2 \cdot \|\mathbf{v}\|_2)\} \in [0, \pi]$. For a vector $\mathbf{w} \in \mathbb{R}^p$ and a subspace $\mathcal{V}$ of $\mathbb{R}^p$, we write $P_{\mathcal{V}}$ for the orthogonal projection matrix onto $\mathcal{V}$, and $P_{\mathcal{V}}\mathbf{w}$ for the orthogonal projection of $\mathbf{w}$ onto $\mathcal{V}$. We write the angle between $\mathbf{w}$ and a subspace $\mathcal{V}$ as $\text{Angle}(\mathbf{w}, \mathcal{V}) = \arccos\{\mathbf{w}'P_{\mathcal{V}}\mathbf{w}/(\|\mathbf{w}\|_2 \cdot \|P_{\mathcal{V}}\mathbf{w}\|_2)\}$. The discrepancy between two subspaces $\mathcal{V}$ and $\tilde{\mathcal{V}}$ can be measured by several canoncial angles (Stewart & Sun, 1990; Jung et al., 2012), and we use the largest canonical angle. Let $\mathbf{V}$ and $\tilde{\mathbf{V}}$ be any orthonormal bases for $\mathcal{V}$ and $\tilde{\mathcal{V}}$, respectively. The largest canonical angle is $\text{Angle}(\mathcal{V}, \tilde{\mathcal{V}}) = \arccos(\gamma_1)$, where γ_1 is the smallest singular value of $\mathbf{V}'\tilde{\mathbf{V}}$.

Assumption 3 (Eigenspace Mean Difference Separability) For each $1 \le i \le m$ and $1 \le j \le K$, $\text{Angle}(\mathbf{u}_i, \boldsymbol{\mu}_j)$ has a limit as $p \to \infty$. Moreover, there exists $\epsilon > 0$ such that $\lim_{p\to\infty} \text{Angle}(\boldsymbol{\mu}_{ij}, P_{\mathcal{U}_m}\boldsymbol{\mu}_{ij}) > \epsilon$ for all $1 \le i < j \le K$.

Assumption 3 extends Assumption 3 of Chang et al. (2021) to a multi-category setting. The first part, regarding the existence of limiting angles, is mostly due to technical reasons and appears unavoidable. The second part mandates that the pairwise population mean difference vectors form non-degenerate angles with the leading eigenspace. Put differently, as per Assumption 1, $p^{-1}\boldsymbol{\mu}_{ij}' P_{\mathcal{U}_m^\perp}\boldsymbol{\mu}_{ij}$ converges to a positive quantity. Here, $\mathcal{U}_m^\perp$ represents the orthogonal complement of $\mathcal{U}_m$, signifying the subspace with relatively smaller noise. The limit of the quantity $p^{-1}\boldsymbol{\mu}_{ij}' P_{\mathcal{U}_m^\perp}\boldsymbol{\mu}_{ij}$ represents the (scaled) projected population mean difference and plays a key role in distinguishing two different classes when projected onto the proposed discriminant

hyperplane (see Lemma 3). Therefore, Assumption 3 may not be necessary to observe the second data piling phenomenon; rather, it is intended to rule out the cases where two distinct classes are being piled to the same point.

Finally, we control the dependency among the true principal component scores in $\mathbf{z}_{kj} = \mathbf{\Lambda}^{-1/2}\mathbf{U}'(X_{kj} - \boldsymbol{\mu}_k) \in \mathbb{R}^p$ by the ρ-mixing condition (Jung & Marron, 2009; Chang et al., 2021). The ρ-mixing condition enables us to weaken the normality assumption for the data or independence assumption on the principal component scores. For detailed explanations of the ρ-mixing condition, see Kolmogorov and Rozanov (1960).

Assumption 4 The elements of the p-vector $\mathbf{z}_{kj}$ have uniformly bounded fourth moments, and for each p, $\mathbf{z}_{kj}$ consists of the first p elements of an infinite random sequence $(z_{k1}, z_{k2}, ...)_j$, which is ρ-mixing under some permutation.

2.2 Second data piling for binary classification

In this subsection, we temporarily assume $K = 2$ and review the *second data piling* (SDP) phenomenon occurring in a binary classification problem. The term 'second' data piling was introduced in Chang et al. (2021) to distinguish it from the concept of the 'first' data piling. The first data piling represents a finite-sample phenomenon where the training data $\mathcal{X}$ projects onto either of two points along a specified vector, such as MDP defined in (1), whenever $p > n - 2$ (Ahn et al., 2012). In contrast, the SDP refers to a data piling tendency of independent test data $\mathcal{Y}$, which are drawn from the same model as the training data $\mathcal{X}$. Importantly, the SDP is an HDLSS-asymptotic phenomenon where p tends to infinity while n is fixed. We provide a formal definition of the SDP phenomenon for the binary classification setting. For any random observation $X \in \mathbb{R}^p$, we write $\pi(X) = k$ if X comes from the kth class. Recall that $\mathcal{S}_X$ is the subspace spanned by training data (2).

Definition 1 (Second data piling for binary classification (Chang et al., 2021)) We say that a vector $\mathbf{v} \in \mathcal{S}_X$ induces second data piling (SDP) if $p^{-1/2}\mathbf{v}'(Y - Y^*) \xrightarrow{P} 0$ as $p \to \infty$ for any independent observations $Y, Y^* \in \mathcal{Y}$ with $\pi(Y) = \pi(Y^*)$.

Definition 1 implies that $\mathbf{v} \in \mathcal{S}_X$ is a SDP vector if projections of independent test data from the same class onto $\mathbf{v}$ are piled on a same location asymptotically. Chang et al. (2021) showed that a SDP vector can be estimated purely from the training data, and utilizing the estimated SDP vector enables an asymptotic perfect classification. Their findings on the SDP phenomenon are summarized as follows:

- Projections of independent test data $\mathcal{Y}$ onto the subspace $\mathcal{S} = \text{span}(\mathbf{w}_{MDP}) \oplus \text{span}(\{\hat{\mathbf{u}}_i\}_{i=1}^m)$ tend to be respectively distributed along two parallel affine subspaces $\mathcal{T}_1, \mathcal{T}_2$. Hence, any vector $\mathbf{v} \in \mathcal{S}$ yields SDP if $\mathbf{v}$ is asymptotically perpendicular to both $\mathcal{T}_1$ and $\mathcal{T}_2$.

- Let a ridge discriminant vector be defined as $\mathbf{w}_\alpha \propto \alpha_p(\mathbf{S}_W + \alpha_p \mathbf{I}_p)^{-1}(\bar{X}_1 - \bar{X}_2)$, where $\alpha_p = \alpha p$ for a ridge parameter $\alpha \in (-\infty, \infty)$. The ridge discriminant vector projected on $\mathcal{S}$, $\mathbf{v}_\alpha = P_{\mathcal{S}}\mathbf{w}_\alpha$, yields SDP when the ridge parameter α is chosen to be $\alpha = -\tau^2$, resulting in $\mathbf{v}_{-\tau^2}$.
- A linear classification rule utilizing the projected negatively-ridged discriminant vector $\mathbf{v}_{-\tau^2}$ can achieve asymptotic perfect classification.

Extending the above results to the multi-category classification setting is somewhat demanding. For example, the *maximal data piling subspace* $\mathcal{W}_{MDP}$ for multi-category classification may not be a natural extension of $\mathbf{w}_{MDP}$. Also, after reconstructing the signal subspace $\mathcal{S}$ with $\mathcal{W}_{MDP}$ instead of $\mathbf{w}_{MDP}$, we need to confirm whether projecting the ridged linear discriminant subspace onto $\mathcal{S}$ contributes to yield SDP in the context of multi-category classification. Moreover, developing a new linear classification rule that can correctly separate multiple piles of independent test data on an SDP subspace is another difficult problem in multi-category classification.

2.3 Maximal data piling for multi-category classification

We first define the maximal data piling (MDP) subspace for multi-category classification. Let $\mathcal{O}(p,d) := \{\mathbf{V} \in \mathbb{R}^{p\times d} : \mathbf{V}'\mathbf{V} = \mathbf{I}_d\}$. If $p > n - K$, then for any $d = 1, \ldots, K-1$, data piling occurs for any $\mathbf{V} \in \mathcal{O}(p,d)$ satisfying trace $(\mathbf{V}'\mathbf{S}_W\mathbf{V}) = 0$. For $d = 1, \ldots, K-1$, the MDP subspace $\mathcal{W}_d := \mathcal{W}_d(\mathcal{X})$ of dimension d is defined as the subspace spanned by $\tilde{\mathbf{V}}_d$:

$$\tilde{\mathbf{V}}_d \in \underset{\mathbf{V}\in\mathcal{O}(p,d)}{\operatorname{argmax}} \text{ trace } (\mathbf{V}'\mathbf{S}_B\mathbf{V}) \quad \text{subject to} \quad \text{trace } (\mathbf{V}'\mathbf{S}_W\mathbf{V}) = 0. \tag{3}$$

Among the subspaces inducing data piling, the MDP subspace $\mathcal{W}_d$ maximizes the sum of pairwise distances among class-wise sample vectors projected onto the subspace since

$$\text{trace } (\mathbf{V}'\mathbf{S}_B\mathbf{V}) = \frac{1}{2n}\sum_{i=1}^{K}\sum_{j=1}^{K} n_i n_j \left\| P_{\mathcal{V}}\bar{X}_i - P_{\mathcal{V}}\bar{X}_j \right\|_2^2,$$

where $\mathcal{V} = \text{span}(\mathbf{V})$. We give an explicit solution to the constrained maximization problem (3). Let $\mathbf{A}^\dagger$ stand for the Moore–Penrose pseudoinverse of a real-valued matrix $\mathbf{A}$.

Lemma 1 *The solution to* (3) *is the matrix of left singular vectors of* $(\mathbf{I}_p - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\dagger)\mathbf{M} = (\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{M}$, *corresponding to the d largest singular values.*

The matrix $\mathbf{I}_p - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\dagger = \mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W'$ is the projection matrix onto the subspace with zero within-class variance. Since there is no within-class scatter along the projected mean difference subspace span$\{(\mathbf{I}_p - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\dagger)\mathbf{M}\}$, data piling occurs. The reason we consider $d \le K-1$ for the dimension of the MDP subspace $\mathcal{W}_d$ is clear from Lemma 1 as

$(\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W')\mathbf{M}$ is of rank $K-1$ at most. When the n observations in the training data are projected onto $\mathcal{W}_d$, they pile up on exactly K points in $\mathcal{W}_d$, one for each class; all n_k observations from the kth class are projected onto $P_{\mathcal{W}_d}\bar{X}_k = \tilde{\mathbf{V}}_d \tilde{\mathbf{V}}_d' \bar{X}_k$. Note that the solution of (3) obtained by Lemma 1, $\tilde{\mathbf{V}}_d$, is an orthogonal basis of $\mathcal{W}_d$ with a *nestedness* property: For $d_1 \le d$, one can easily check that d_1 columns of $\tilde{\mathbf{V}}_d$ form $\tilde{\mathbf{V}}_{d_1}$. We remark that analyzing $\mathcal{W}_d$ through all pairwise MDP directions $\left\{\mathbf{w}_{MDP,ij} : 1 \le i < j \le K\right\}$, where $\mathbf{w}_{MDP,ij}$ is the MDP direction for the training data from the ith and jth classes, is generally not possible. In Appendix A.1, we include a detailed explanation of the relationship between pairwise MDP directions and the MDP subspace.

In the following sections, we assume that $(\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W')\mathbf{M}$ is of rank $K-1$, and examine the SDP phenomenon in the multi-category classification setting using the $(K-1)$-dimensional MDP subspace $\mathcal{W}_{MDP} := \mathcal{W}_{K-1}$.

3 Second data piling for multi-category classification

In this section, we show that a subspace exhibiting SDP can be obtained purely from the training data $\mathcal{X}$. Moreover, we show that this subspace maximizes the sum of pairwise distances of class-wise independent test data $\mathcal{Y}$ among SDP subspaces. For this, we extend the definition of SDP for binary classification (Definition 1) to the multi-category case.

Definition 2 (Second data piling for multi-category classification) We say that a $(K-1)$-dimensional subspace $\mathcal{V} \subset \mathcal{S}_X$ induces second data piling (SDP) if $p^{-1/2}\left\|P_{\mathcal{V}}(Y - Y^*)\right\|_2 \xrightarrow{P} 0$ as $p \to \infty$ for any independent observations $Y, Y^* \in \mathcal{Y}$ with $\pi(Y) = \pi(Y^*)$. Such a subspace $\mathcal{V}$ is called an SDP subspace.

Definition 2 implies that a $(K-1)$-dimensional subspace $\mathcal{V} \subset \mathcal{S}_X$ is an SDP subspace if independent test data from the same class are asymptotically piled on a same location, one for each class. We write $\mathfrak{C}_{SDP}$ for the collection of $(K-1)$-dimensional subspaces inducing SDP.

3.1 Asymptotic SDP subspaces

We first define the signal subspace which captures a strong variation in the data for multi-category classification. Note that $\mathcal{W}_{MDP}$ is orthogonal to the sample eigenvectors of $\mathbf{S}_W$, and we can decompose the sample space $\mathcal{S}_X$ into $\mathcal{S}_X = \mathcal{W}_{MDP} \oplus \text{span}(\{\hat{\mathbf{u}}_i\}_{i=1}^{n-K})$. Among the sample eigenvectors of $\mathbf{S}_W$, the leading sample eigenvectors $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m$ are capable of capturing important variability since the true leading eigenspace $\mathcal{U}_m = \text{span}(\{\mathbf{u}_i\}_{i=1}^{m})$ can be (partially) estimated by $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m$ (the limiting distribution of $\text{Angle}(\hat{\mathbf{u}}_i, \mathbf{u}_j)$, for $i, j = 1, \ldots, m$, is non-degenerate). In contrast, the other sample eigenvectors $\hat{\mathbf{u}}_{m+1}, \ldots, \hat{\mathbf{u}}_{n-K}$ are *strongly inconsistent* with the leading eigenvectors in the sense that $\text{Angle}(\hat{\mathbf{u}}_i, \mathbf{u}_j) \xrightarrow{P} \pi/2$ for

$i = m+1, \ldots, n-K$ and $j = 1, \ldots, m$; see Appendix A.2 and (Jung et al., 2012). Removing the unimportant subspace $\mathcal{S}^{\perp} = \text{span}(\{\hat{\mathbf{u}}_i\}_{i=m+1}^{n-K})$ from $\mathcal{S}_X$, we define $\mathcal{S} = \mathcal{W}_{MDP} \oplus \text{span}(\{\hat{\mathbf{u}}_i\}_{i=1}^{m})$ and call $\mathcal{S}$ the signal subspace of the sample space $\mathcal{S}_X$.

Within the signal subspace $\mathcal{S}$, we show that a subspace orthogonal to the leading eigenspace $\mathcal{U}_m = \text{span}(\{\mathbf{u}_i\}_{i=1}^{m})$ yields SDP. For this, let $\mathcal{T} = \text{span}(\{\mathbf{u}_{i,\mathcal{S}}\}_{i=1}^{m})$ where $\mathbf{u}_{i,\mathcal{S}} = P_{\mathcal{S}}\mathbf{u}_i$ for $i = 1, \ldots, m$. Decompose $\mathcal{S}$ into the two subspaces $\mathcal{T}$ (of dimension m) and $\mathcal{H} := \mathcal{T}^{\perp}|_{\mathcal{S}} = \mathcal{T}^{\perp} \cap \mathcal{S}$ (of dimension $K-1$), so that $\mathcal{S} = \mathcal{T} \oplus \mathcal{H}$.

Theorem 1 *Suppose that Assumptions 1–4 hold. Then for any $Y \in \mathcal{Y}$ with $\pi(Y) = k$,*

$$p^{-1/2} \left\| P_{\mathcal{H}}(Y - \boldsymbol{\mu}_k) \right\|_2 \xrightarrow{P} 0 \tag{4}$$

as $p \to \infty$ for all $k = 1, \ldots, K$. Hence, $\mathcal{H}$ is an SDP subspace, i.e., $\mathcal{H} \in \mathfrak{G}_{SDP}$.

Theorem 1 implies that projections of independent test data $\mathcal{Y}$ from the kth class onto $\mathcal{S}$ tend to lie on an m-dimensional affine subspace $\mathcal{T}_k \subset \mathcal{S}$ given by

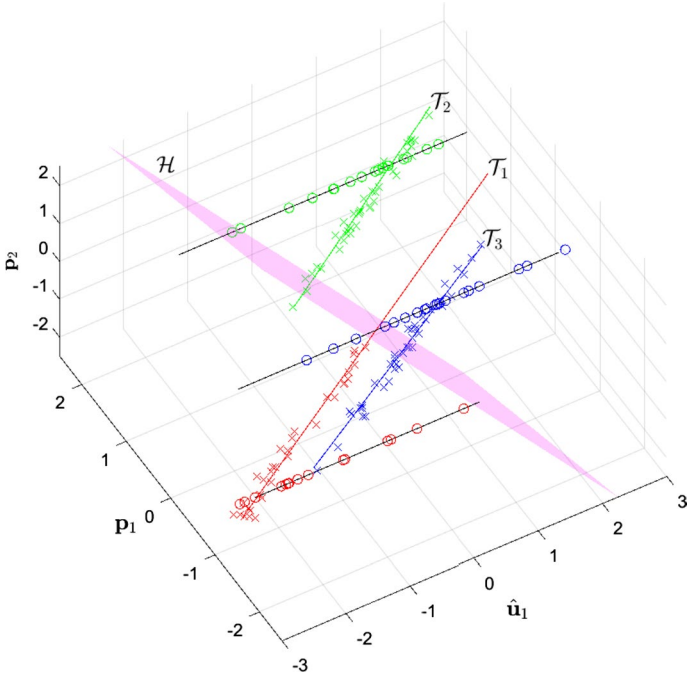


Fig. 3 Projections of training data (circles) and independent test data (crosses) onto $\mathcal{S} = \mathcal{W}_{MDP} \oplus \text{span}(\hat{\mathbf{u}}_1)$, where $\mathcal{W}_{MDP} = \text{span}(\mathbf{p}_1, \mathbf{p}_2)$. Colors code three different classes. In this case, $\mathcal{H}$ is 2-dimensional subspace which is the orthogonal complement of parallel lines $\mathcal{T}_k$ $(k = 1, 2, 3)$ within 3-dimensional subspace $\mathcal{S} = \mathcal{W}_{MDP} \oplus \text{span}(\hat{\mathbf{u}}_1)$

$$\mathcal{T}_k = \left\{ \mathbf{v} + (P_{\mathcal{S}} - P_{\mathcal{T}})\boldsymbol{\mu}_k \,:\, \mathbf{v} \in \mathcal{T} \subset \mathbb{R}^p \right\}. \tag{5}$$

Note that the m-dimensional affine subspace $\mathcal{T}_k$ is parallel to the subspace $\mathcal{T}$ and to each other. Figure 3 demonstrates the SDP phenomenon using a single spiked covariance model with $m = 1$. While the training data are piled on exactly $K = 3$ points on $\mathcal{W}_{MDP}$, independent test data are not. Instead, they are concentrated along parallel lines $\mathcal{T}_k$ $(k = 1, 2, 3)$ within $\mathcal{S} = \mathcal{W}_{MDP} \oplus \text{span}(\hat{\mathbf{u}}_1)$, one for each class. These parallel lines are orthogonal to the SDP subspace $\mathcal{H}$, and projections of independent test data onto $\mathcal{H}$ will be piled on top of each other, one for each class.

One may wonder whether there are other subspaces within the sample space $\mathcal{S}_X = \mathcal{S} \oplus \mathcal{S}^\perp$ that can yield SDP of independent test data. In fact, there are infinitely many such subspaces (Lemma 2), but $\mathcal{H} \subset \mathcal{S}$ is *maximal* in the sense that when independent test data are projected onto the subspace, it induces the maximal separation among the class-wise projections of independent test data (Theorem 2).

Lemma 2 *Suppose that Assumptions 1–4 hold. Then* $\mathcal{V} \in \mathfrak{C}_{SDP}$ *if and only if* $\left\|P_{\mathcal{V}}\mathbf{u}_{i,\mathcal{S}}\right\|_2 \xrightarrow{P} 0$ *for all* $i = 1, \ldots, m$ *as* $p \to \infty$. *Moreover, for any given* $\mathcal{V} \in \mathfrak{C}_{SDP}$, *there exists* $\tilde{\mathcal{V}} \in \mathcal{B}$ *such that* $\text{Angle}(\mathcal{V}, \tilde{\mathcal{V}}) \xrightarrow{P} 0$ *as* $p \to \infty$ *where* $\mathcal{B} = \{\tilde{\mathcal{V}} \subset \mathcal{S}_X \,:\, \dim(\tilde{\mathcal{V}}) = K - 1, \tilde{\mathcal{V}} \subset \mathcal{H} \oplus \mathcal{S}^\perp\}$.

Lemma 2 states that the key condition for a subspace $\mathcal{V} \subset \mathcal{S}_X$ to achieve SDP: Asymptotic orthogonality to the leading eigenspace. Moreover, any SDP subspace lies in $\mathcal{H} \oplus \mathcal{S}^\perp$. Since $P_{\mathcal{S}^\perp}\mathbf{u}_i \approx \mathbf{0}_p$ for sufficiently large p for all $i = 1, \ldots, m$, any subspace in $\mathcal{S}^\perp$ induces SDP.

To show the maximality of $\mathcal{H}$ among SDP subspaces, we formally define an asymptotic distance between two piles of independent test data from two different classes. For $\mathcal{V} \in \mathfrak{C}_{SDP}$, let $D_{ij}(\mathcal{V})$ be the probability limit of $p^{-1/2}\left\|P_{\mathcal{V}}(Y_i - Y_j)\right\|_2$ as $p \to \infty$, if it exists for any independent test data $Y_i, Y_j \in \mathcal{Y}$ with $\pi(Y_i) = i$ and $\pi(Y_j) = j, 1 \le i \ne j \le K$.

Theorem 2 *Suppose that Assumptions 1–4 hold. Then for any* $i \ne j$ *and* $\mathcal{V} \in \mathfrak{C}_{SDP}$ *such that* $D_{ij}(\mathcal{V})$ *exists,* $D_{ij}(\mathcal{V}) \le D_{ij}(\mathcal{H})$ *with probability* 1.

Theorem 2 implies that the SDP subspace $\mathcal{H}$ maximizes the sum of pairwise distances among different classes of projected test data, and in this sense, $\mathcal{H}$ may be called the maximal SDP subspace.

3.2 Estimation of the maximal SDP subspace

In this subsection, we show that estimation of the maximal SDP subspace $\mathcal{H} = \mathcal{T}^\perp|_{\mathcal{S}}$ is possible. Recent researches have shown that the optimal ridge parameter can be negative in the context of linear regression (Kobak et al., 2020; Tsigler & Bartlett, 2020; Wu & Xu, 2020) and binary classification (Chang et al., 2021) to compensate

an implicit regularization in the overparametrized regime. For multi-category classification, we show that a negatively ridged discriminant subspace can yield SDP when projected onto the signal subspace $\mathcal{S}$. For a given ridge parameter $\alpha \in \mathbb{R}$, let

$$\mathbf{L}_{\alpha,i} = \alpha_p(\mathbf{S}_W + \alpha_p \mathbf{I}_p)^{-1}\mathbf{d}_i \tag{6}$$

where $\alpha_p = \alpha p$ and $\mathbf{d}_i = \bar{X}_i - \bar{X}_K$ for $i = 1, \ldots, K-1$. We note that the first MDP subspace $\mathcal{W}_{MDP}$ is a limiting point of the ridged discriminant subspace

$$\mathcal{L}_\alpha = \text{span}(\{\mathbf{L}_{\alpha,i}\}_{i=1}^{K-1}) \tag{7}$$

when $\alpha \to 0$, since $\lim_{\alpha\to 0} \mathbf{L}_{\alpha,i} = (\mathbf{I}_p - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\dagger)\mathbf{d}_i$ and $\text{span}(\mathbf{D}) = \text{span}(\mathbf{M})$ where $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{K-1}]$. We define the projected ridged discriminant subspace to be

$$\mathcal{H}_\alpha = \text{span}(\{P_{\mathcal{S}}\mathbf{L}_{\alpha,i}\}_{i=1}^{K-1}), \tag{8}$$

that is, $\mathcal{H}_\alpha$ is the subspace generated by projections of $\mathbf{L}_{\alpha,i}$ $(i = 1, \ldots, K-1)$ onto $\mathcal{S}$. The following theorem shows that for a careful choice of α, the subspace $\mathcal{H}_\alpha$ approximates the maximal SDP subspace $\mathcal{H}$ (asympotically) and it occurs when the ridge parameter α is set as a consistent estimate $\hat{\alpha}$ of the negative ridge parameter $-\tau^2$. Recall that τ^2 is the asymptotic average of the non-leading eigenvalues of the common covariance matrix $\mathbf{\Sigma}$ as mentioned in Assumption 2. Note that $\hat{\alpha}$ can be obtained purely from the training data since $p^{-1}\hat{\lambda}_i \xrightarrow{P} \tau^2$ as $p \to \infty$ for $i = m+1, \ldots, n-K$ (Jung et al., 2012). From now on, we fix

$$\hat{\alpha} = -\frac{1}{n-m-K}\sum_{i=m+1}^{n-K}\frac{\hat{\lambda}_i}{p}. \tag{9}$$

Theorem 3 *Suppose that Assumptions 1—4 hold. Then* (i) $\text{Angle}(\mathcal{H}_{\hat{\alpha}}, \mathcal{T}) \xrightarrow{P} \pi/2$ *as* $p \to \infty$. *Hence,* $\mathcal{H}_{\hat{\alpha}}$ *is an SDP subspace, i.e.,* $\mathcal{H}_{\hat{\alpha}} \in \mathfrak{C}_{SDP}$. (ii) *The subspace* $\mathcal{H}_{\hat{\alpha}}$ *is a consistent estimator of the maximal SDP subspace in the sense that* $\text{Angle}(\mathcal{H}_{\hat{\alpha}}, \mathcal{H}) \xrightarrow{P} 0$ *as* $p \to \infty$.

The use of the estimated SDP subspace $\mathcal{H}_{\hat{\alpha}}$ is demonstrated in Fig. 1c, in which projections of independent test data onto $\mathcal{H}_{\hat{\alpha}}$ are (asymptotically) piled on distinct points.

We remark that if $m = 0$, then the maximal SDP subspace $\mathcal{H} = \mathcal{T}^\perp|_{\mathcal{S}}$ naturally becomes $\mathcal{S} = \mathcal{W}_{MDP}$. That is, when the variables are weakly correlated, $\mathcal{W}_{MDP}$ yields both of the first data piling and SDP.

4 Bias-corrected nearest centroid classifier on a projected ridge subspace

In this section, we provide a new multi-category classification rule using the estimated SDP subspace $\mathcal{H}_{\hat{\alpha}}$, defined in Sect. 3.2, and show that it achieves asymptotic perfect classification. This classification rule is modified from the well-known nearest centroid classification rule, which assigns new independent observation Y to the label of the class of the training data whose centroid is nearest to Y. We show that a typical nearest centroid classification rule may not achieve successful classification performances due to a bias, but this bias can be adjusted and thus asymptotic perfect classification is possible. Let $\boldsymbol{\delta}(k,j)$ be a $(K-1)$-vector whose ith coordinate is $\lim_{p\to\infty} p^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_K)' P_{\mathcal{U}_m^\perp}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)$, where $\mathcal{U}_m^\perp = \text{span}(\{\mathbf{u}_i\}_{i=m+1}^p)$, the orthogonal complement of the leading eigenspace. That is, $\boldsymbol{\delta}(k,j)$ represents the limit of scaled inner products between the expectation of the matrix of the sample mean differences $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{K-1}]$ (recall that $\mathbf{d}_i = \bar{X}_i - \bar{X}_K$ for $i = 1, \ldots, K-1$) and the true mean difference of the kth and jth classes when they are projected onto $\mathcal{U}_m^\perp$. Lemma 3 shows that the empirical mean differences projected onto the estimated SDP subspace $\mathcal{H}_{\hat{\alpha}}$ are asymptotically decomposed into the sum of the 'true' mean differences and a bias term. We write $\mathbf{H}_\alpha = [\boldsymbol{\omega}_{\alpha,1}, \ldots, \boldsymbol{\omega}_{\alpha,K-1}]$ where $\boldsymbol{\omega}_{\alpha,i} = p^{-1/2} P_{\mathcal{S}} \mathbf{L}_{\alpha,i}$ for $i = 1, \ldots, K-1$ so that $\mathcal{H}_\alpha = \text{span}(\mathbf{H}_\alpha)$. Recall that $\mathcal{H}_\alpha$ is the projected ridged discriminant subspace defined in (8).

Lemma 3 *Suppose that Assumptions 1—4 hold. For any independent observation $Y \in \mathcal{Y}$ and $j = 1, \ldots, K$,*

$$p^{-1/2}\mathbf{H}_{\hat{\alpha}}'(Y - \bar{X}_j) \xrightarrow{P} \boldsymbol{\delta}(\pi(Y), j) + \boldsymbol{\xi}_j, \tag{10}$$

as $p \to \infty$ where $\boldsymbol{\xi}_j = -n_j^{-1}\tau^2 \mathbf{e}_j$ if $j = 1, \ldots, K-1$ and $\boldsymbol{\xi}_j = n_j^{-1}\tau^2 \mathbf{1}$ if $j = K$. Here, $\mathbf{e}_j \in \mathbb{R}^{K-1}$ is the jth standard unit vector in $\mathbb{R}^{K-1}$, and $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^{K-1}$.

We remark that if $k = j$, then $\boldsymbol{\delta}(k,j) = \mathbf{0}_{K-1}$. In contrast, if $k \neq j$, then $\boldsymbol{\delta}(k,j) \neq \mathbf{0}_{K-1}$ since Assumption 3 implies $\text{Angle}(\boldsymbol{\mu}_{kj}, P_{\mathcal{U}_m^\perp}\boldsymbol{\mu}_{kj}) \nrightarrow \pi/2$ as $p \to \infty$. Heuristically, the asymptotic distance between projections of the mean of training data of the jth class $\bar{X}_j$ and independent test data Y onto $\mathcal{H}_{\hat{\alpha}}$ can be completely decomposed into the projection of true mean differences $\boldsymbol{\delta}(\pi(Y), j)$ and the bias term ξ_j. The bias term ξ_j does not depend on the class of Y, but inevitably exists due to the accumulated noise under the HDLSS asymptotic regime. Nevertheless, the *bias* $\boldsymbol{\xi}_j$ can be consistently estimated by $\hat{\boldsymbol{\xi}}_j$, obtained by replacing τ^2 in $\boldsymbol{\xi}_j$ with $-\hat{\alpha}$. Subtract $\hat{\boldsymbol{\xi}}_j$ from the projections (10), we obtain a bias-corrected (coordinate-wise) signed distances to the jth class

$$\mathbf{g}_{\hat{\alpha},j}(Y;\mathcal{X}) = p^{-1/2}\mathbf{H}_{\hat{\alpha}}'(Y - \bar{X}_j) - \hat{\boldsymbol{\xi}}_j, \tag{11}$$

for $j = 1, \ldots, K$. Based on the above construction, we define a bias-corrected nearest centroid classification rule $\phi_{PRS-BCNC,\alpha}$ using (11), for given α, by

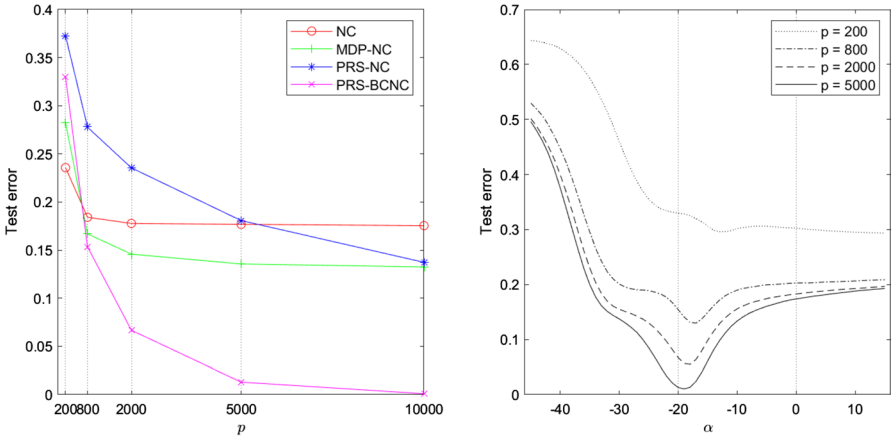
$$\phi_{PRS-BCNC,\alpha}(Y;\mathcal{X}) = \underset{j=1,\ldots,K}{\operatorname{argmin}} \left\| \mathbf{g}_{\alpha,j}(Y;\mathcal{X}) \right\|_2. \tag{12}$$

This classification rule utilizes the fact that the distance between new independent observation and the centroid can be estimated on the subspace $\mathcal{H}_{\hat{\alpha}}$. The following theorem states that $\phi_{PRS-BCNC,\alpha}$ can achieve asymptotic perfect classification when we use the negative ridge parameter $\alpha = \hat{\alpha} < 0$.

Theorem 4 *Suppose that Assumptions 1—4 hold. Then* $\lim_{p\to\infty} \mathbb{P}\{\phi_{PRS-BCNC,\hat{\alpha}}(Y;\mathcal{X}) \neq \pi(Y)\} = 0$.

5 Numerical studies

We conduct numerical studies to show that $\phi_{PRS-BCNC,\alpha}$ achieves asymptotic perfect multi-category classification when we use the negative ridge parameter $\hat{\alpha}$, which is a consistent estimate of $-\tau^2$. First, we provide a simulation experiment result. We assume $K = 3$, and for $i = 1, 2, 3$ and $j = 1, \ldots, 20$, let $X_{ij} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_1 = (\mathbf{1}_{p/2}, \sqrt{3}\mathbf{1}_{p/2})'$, $\boldsymbol{\mu}_2 = (\sqrt{3}\mathbf{1}_{p/2}, -\mathbf{1}_{p/2})'$, $\boldsymbol{\mu}_3 = \mathbf{0}_p$, and $\boldsymbol{\Sigma} = 20\mathbf{I}_p + \mathbf{1}_p\mathbf{1}_p'$. We compare $\phi_{PRS-BCNC,\alpha}$ with ϕ_{NC} (nearest centroid classification rule), ϕ_{MDP-NC} (nearest centroid classification rule on $\mathcal{W}_{MDP}$), and $\phi_{PRS-NC,\alpha}$ (nearest centroid classification rule on the projected ridge subspace $\mathcal{H}_\alpha$). Using these classification rules, we obtain classification error rates using 1,500 independent observations (500 independent observations for each class). This procedure is repeated for 100 times and we report the average of the repetitions. Figure 4a shows that $\phi_{PRS-BCNC,\alpha}$ achieves



(a) (b)

Fig. 4 **a** The classification error rates of ϕ_{NC}, ϕ_{MDP-NC}, $\phi_{PRS-NC,\hat{\alpha}}$ and $\phi_{PRS-BCNC,\hat{\alpha}}$ for $p = 200, 800, 2000, 5000, 10000$. **b** The classification error rates of $\phi_{PRS-BCNC,\alpha}$ over a fine grid of $\alpha \in [-45, 15]$

Table 1 Classification error bars (95% confidence intervals) of the four classification rules

p	ϕ_{NC}	ϕ_{MDP-NC}	$\phi_{PRS-NC,\hat{\alpha}}$	$\phi_{PRS-BCNC,\hat{\alpha}}$
200	0.236 ± 0.031	0.282 ± 0.040	0.372 ± 0.061	0.330 ± 0.060
800	0.184 ± 0.035	0.167 ± 0.022	0.278 ± 0.063	0.154 ± 0.041
2000	0.178 ± 0.034	0.146 ± 0.022	0.236 ± 0.066	0.067 ± 0.031
5000	0.177 ± 0.040	0.136 ± 0.023	0.181 ± 0.052	0.013 ± 0.010
10000	0.176 ± 0.027	0.133 ± 0.024	0.137 ± 0.077	0.001 ± 0.001

nearly perfect classification for sufficiently large p when we use $\alpha := \hat{\alpha} < 0$ while the other classification rules do not. This result implies that both of the SDP phenomenon occurring on $\mathcal{H}_{\hat{\alpha}}$ and the bias-correction strategy contribute to achieving perfect classification. We also check the classification error of $\phi_{PRS-BCNC,\alpha}$ over a fine grid of $\alpha \in [-45, 15]$. Figure 4b indicates that $\phi_{PRS-BCNC,\alpha}$ achieves its minimum (and nearly zero) classification error with the negative ridge parameter $\alpha = -\tau^2 = -20 < 0$ when p is sufficiently large, which justifies that the optimal ridge parameter can be negative in high-dimensional classification. Table 1 displays classification error bars of ϕ_{NC}, ϕ_{MDP-NC}, $\phi_{PRS-NC,\hat{\alpha}}$ and $\phi_{PRS-BCNC,\hat{\alpha}}$ for the simulation study. This result clearly shows that only $\phi_{PRS-BCNC,\hat{\alpha}}$ achieves nearly perfect classification as the dimension p increases among the four classification rules. Nearly identical conclusions were obtained under a two-component spiked covariance model and a four-category classification situation $K = 4$.

Moreover, we show that the negative ridge phenomenon can also be observed with $\phi_{PRS-BCNC,\alpha}$ using random Fourier features on MNIST dataset (LeCun et al., 2010). After normalizing the 784 pixel intensity values of each image to $[-1, 1]$, we calculate $\exp(-i\mathbf{X}'\mathbf{W})$ where $\mathbf{W} \in \mathbb{R}^{784\times 2000}$ is a random matrix whose each element is independent and identically distributed from $\mathcal{N}(0, 0.1^2)$, as done in Chang et al. (2021). We regard the real and imaginary parts as separate variables so that the feature space has dimension $p = 4,000$. We convert the original ten-category classification problem to ${}_{10}C_3 = 120$ three-category classification problems. For each class,

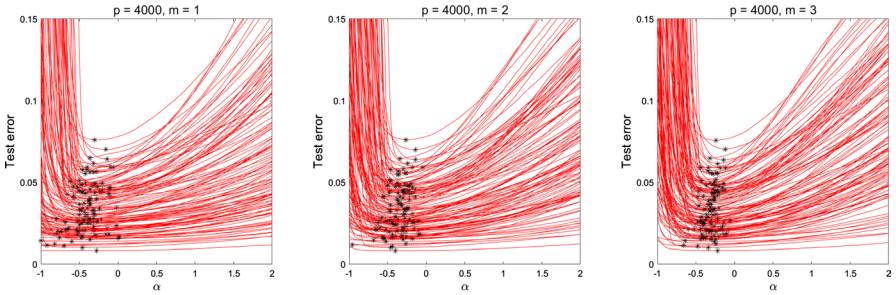


Fig. 5 The classification error rates of $\phi_{PRS-BCNC,\alpha}$ over a fine grid of $\alpha \in [-1, 2]$ for three-category classification problems of the MNIST dataset using random Fourier features with $m = 1, 2, 3$. The minimum classification error rate of each problem is marked as a star

we randomly choose 100 images as training data, and the remaining images are used as test data. The classification error of $\phi_{PRS-BCNC,\alpha}$ is evaluated using the test data over a fine grid of $\alpha \in [-1, 2]$. We repeat this procedure ten times and report the average of the repetitions. Figure 5 shows that, when we assume $m = 1, 2, 3$, the minimum classification error is achieved when negative ridge parameters are used for nearly all classification problems.

6 Conclusion

The double data piling phenomenon uncovered in this study refers to two distinct phenomena: The first data piling, observed in finite-dimensional cases during the training phase, and the second data piling, which occurs asymptotically when the dimension tends to infinity during the testing phase. While existing research on double data piling has focused on binary classification problems, its applicability is limited when it comes to real-world classification scenarios involving multiple categories or classes. We address this limitation by providing a comprehensive characterization of the asymptotic perfect classification of high-dimensional data, encompassing multi-category classification problems.

As the present work assumes homogeneous covariance for each class, further generalization can be pursued in various directions of heterogeneous covariance structures: different numbers of spikes, varying leading eigenvalues and/or eigenvectors, or the size of the noise components. In fact, our claims in this article can be further extended to heterogeneous spiked covariance models (Kim et al., 2022). For the ith class covariance matrix $\mathbf{\Sigma}_{(i)}$, write the eigen-decomposition $\mathbf{\Sigma}_{(i)} = \mathbf{U}_{(i)}\mathbf{\Lambda}_{(i)}\mathbf{U}'_{(i)}$ where $\mathbf{\Lambda}_{(i)} = \text{diag}\,(\{\lambda_{(i),j}\}_{j=1}^{p})$ arranged in descending order and $\mathbf{U}_{(i)} = [\mathbf{u}_{(i),1}, \ldots, \mathbf{u}_{(i),p}]$. For each class (i.e., each i), assume the first m_i eigenvalues increase at the order of p, that is, $\lambda_{(i),j} = \sigma^2_{(i),j}p$ for $j = 1, \ldots, m_i$, while $p^{-1}\sum_{j=m_i+1}^{p}\lambda_{(i),j} \to \tau^2 \in (0, \infty)$ as $p \to \infty$. Define the common leading eigenspace $\mathcal{U}_m = \text{span}(\{\mathcal{U}_{(i)}\}_{i=1}^{K})$ where $\mathcal{U}_{(i)} = \text{span}(\{\mathbf{u}_{(i),j}\}_{j=1}^{m_i})$ and assume that the rank of $\mathcal{U}_m$ is m. Then even if the number of spikes m_i or the leading eigenspace of the ith class $\mathcal{U}_{(i)}$ are different for all $i = 1, \ldots, K$, the negatively projected ridged subspace $\mathcal{H}_{\hat{\alpha}}$ approximates the maximal second data piling subspace, and our proposed method utilizing $\mathcal{H}_{\hat{\alpha}}$ achieves asymptotic multi-category perfect classification. However, when $p^{-1}\sum_{j=m_i}^{p}\lambda_{(i),j} \to \tau^2_{(i)}$ as $p \to \infty$ and $\tau^2_{(i)} \neq \tau^2_{(j)}$ for $i \neq j$, our proposed strategy does not work successfully since $\mathcal{H}_{\hat{\alpha}}$ does not yield second data piling. Nevertheless, we expect that a second data pilling subspace can be estimated in such general settings, and thus asymptotic perfect classification is possible. A comprehensive investigation of this matter remains a topic for future research.

Technical details and proofs

Relationship between pairwise MDP directions and the MDP subspace

Suppose that we take any two classes (say, the ith and the jth classes), and examine the pairwise MDP direction for the subset of the training data corresponding to the $\{i,j\}$ pair of classes. The within-scatter matrix for the $\{i,j\}$ pair is $\mathbf{S}_{\{i,j\}} = \tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i' + \tilde{\mathbf{X}}_j\tilde{\mathbf{X}}_j'$, whose the orthogonal complement subspace, denoted by $\mathcal{V}_{\{i,j\}}^{\perp}$, is of dimension $p-(n_i+n_j-2)$ with probability one. In this case, the MDP direction is obtained by the orthogonal projection of $\bar{X}_i - \bar{X}_j$ onto $\mathcal{V}_{\{i,j\}}^{\perp}$:

$$\mathbf{w}_{MDP,ij} := P_{\mathcal{V}_{\{i,j\}}^{\perp}}(\bar{X}_i - \bar{X}_j) / \left\| P_{\mathcal{V}_{\{i,j\}}^{\perp}}(\bar{X}_i - \bar{X}_j) \right\|_2.$$

Is $\mathbf{w}_{MDP,ij}$ contained in the MDP subspace $\mathcal{W}_{MDP}$? If the answer to this question is true, then one may try analyzing $\mathcal{W}_{MDP}$ through all pairwise MDP directions $\{\mathbf{w}_{MDP,ij} : 1 \le i < j \le K\}$. We show in the following that this is not the case, since $\mathbf{w}_{MDP,ij} \in \mathcal{W}_{MDP}$ is generally not true. Fix i and j, and decompose the within-class scatter matrix by

$$\mathbf{S}_W = \mathbf{S}_{\{i,j\}} + \sum_{k \notin \{i,j\}} \tilde{\mathbf{X}}_k\tilde{\mathbf{X}}_k', \tag{A1}$$

and write $\hat{\mathbf{U}}_W$ and $\hat{\mathbf{U}}_{\{i,j\}}$ be the orthogonal basis matrices for $\mathbf{S}_W$ and $\mathbf{S}_{\{i,j\}}$, respectively. Let $\hat{\mathbf{U}}_{\{i,j\}^c}$ be a full-rank orthogonal matrix satisfying

$$\hat{\mathbf{U}}_W\hat{\mathbf{U}}_W' = \hat{\mathbf{U}}_{\{i,j\}}\hat{\mathbf{U}}_{\{i,j\}}' + \hat{\mathbf{U}}_{\{i,j\}^c}\hat{\mathbf{U}}_{\{i,j\}^c}', \tag{A2}$$

and $\hat{\mathbf{U}}_{\{i,j\}}'\hat{\mathbf{U}}_{\{i,j\}^c} = \mathbf{0}$. We observe the following.

Proposition 1 *Assume that the data are in general position, that is, no three points are collinear, and $K \ge 3$, $n_k \ge 2$ for all $k = 1, \ldots, K$. Then for any $1 \le i < j \le K$, the pairwise MDP direction is not contained in the MDP subspace, (i.e., $\mathbf{w}_{MDP,ij} \notin \mathcal{W}_{MDP}$) if and only if $\hat{\mathbf{U}}_{\{i,j\}^c}'(\bar{X}_i - \bar{X}_j) \ne \mathbf{0}$.*

Proof of Proposition 1 The assumptions of general position and $K \ge 3$, $n_k \ge 2$ guarantee that $\hat{\mathbf{U}}_{\{i,j\}^c}$ has more than one column. By (A2), we have, for a normalizing constant $c > 0$,

$$\begin{aligned}
\mathbf{w}_{MDP,ij} &= c(\mathbf{I}_p - \hat{\mathbf{U}}_{\{i,j\}}\hat{\mathbf{U}}_{\{i,j\}}')(\bar{X}_i - \bar{X}_j) \\
&= c(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')(\bar{X}_i - \bar{X}_j) + c\hat{\mathbf{U}}_{\{i,j\}^c}\hat{\mathbf{U}}_{\{i,j\}^c}'(\bar{X}_i - \bar{X}_j) \\
&=: \mathbf{w}_1 + \mathbf{w}_2.
\end{aligned}$$

Recall that $\mathcal{W}_{MDP} = \mathrm{span}\{(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{M}\}$. Let $\mathbf{b}_i = (0, \ldots, 0, {n_i}^{-1/2}, 0, \ldots, 0)' \in \mathbb{R}^K$ whose ith element is nonzero. Since $\mathbf{w}_1 = c(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{M}(\mathbf{b}_i - \mathbf{b}_j)$, $\mathbf{w}_1 \in \mathcal{W}_{MDP}$.

However, since $(\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W')\hat{\mathbf{U}}_{\{i,j\}^c} = \mathbf{0}$ again by (A2), $\mathbf{w}_2 \notin \mathcal{W}_{MDP}$. Thus $\mathbf{w}_{MDP,ij} \in \mathcal{W}_{MDP}$ if and only if $\|\mathbf{w}_2\|_2 = 0$, or if $\hat{\mathbf{U}}_{\{i,j\}^c}'(\bar{X}_i - \bar{X}_j) = \mathbf{0}$.

Note that the condition $\hat{\mathbf{U}}_{\{i,j\}^c}'(\bar{X}_i - \bar{X}_j) \neq \mathbf{0}$ holds generically. For example, if the data for each class are sampled from a continuous distribution with a full-rank covariance matrix, the condition holds with probability one.

Asymptotic properties of high-dimensional sample within-scatter matrix

In this section, we focus on asymptotic properties of eigenvalues and eigenvectors of the sample within-scatter matrix $\mathbf{S}_W = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$. Note that the centered data matrix $\tilde{\mathbf{X}}$ can be expressed as $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I}_n - \mathbf{J})$ where $\mathbf{X}$ is the $p \times n$ data matrix and $\mathbf{J}$ is an $n \times n$ block diagonal matrix,

$$\mathbf{J} = \operatorname{diag}\left(\frac{1}{n_i}\mathbf{1}_{n_i}\mathbf{1}_{n_i}'\right)_{i=1,\ldots,K}.$$

Here, $\mathbf{1}_l$ is an l-dimensional vector with ones. Recall that we write an eigen-decomposition of $\mathbf{S}_W$ as $\mathbf{S}_W = \hat{\mathbf{U}}_W \hat{\mathbf{\Lambda}}_W \hat{\mathbf{U}}_W'$. Also, the eigen-decomposition of the common covariance matrix $\mathbf{\Sigma}$ is $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$. Let $\mathbf{Z}$ be the standardized principal components of $\mathbf{X}$,

$$\mathbf{Z} = \mathbf{\Lambda}^{-1/2}\mathbf{U}'(\mathbf{X} - \mathbb{E}(\mathbf{X})) = \begin{bmatrix}\mathbf{z}_1' \\ \vdots \\ \mathbf{z}_p'\end{bmatrix} = \begin{bmatrix}\mathbf{z}_{1,1}' & \cdots & \mathbf{z}_{K,1}' \\ \vdots & \vdots & \vdots \\ \mathbf{z}_{1,p}' & \cdots & \mathbf{z}_{K,p}'\end{bmatrix} \in \mathbb{R}^{p \times n}$$

where $\mathbf{z}_{i,j} \in \mathbb{R}^{n_i}$ is the jth principal component scores of the ith class. Then the elements of $\mathbf{Z}$ have mean zero and unit variance and are uncorrelated. We write $\bar{\mathbf{z}}_i = n^{-1}\mathbf{z}_i'\mathbf{1}_n$ and $\bar{\mathbf{z}}_{i,j} = n_i^{-1}\mathbf{z}_{i,j}'\mathbf{1}_{n_i}$. For a square matrix $\mathbf{M}$, we denote $\varphi_i(\mathbf{M})$ and $v_i(\mathbf{M})$ as the ith largest eigenvalue of $\mathbf{M}$ and corresponding eigenvector, respectively. Also, let $v_{ij}(\mathbf{M})$ be the jth coefficient of $v_i(\mathbf{M})$.

Lemma 4 *Suppose that Assumptions 1–4 hold. Let an $n \times m$ matrix of the leading m component scores as $\mathbf{W} = [\sigma_1\mathbf{z}_1, \ldots, \sigma_m\mathbf{z}_m]$ and $\mathbf{\Phi} = \mathbf{W}'(\mathbf{I}_n - \mathbf{J})\mathbf{W}$. Then, the following hold as $p \to \infty$.*

(i) *$p^{-1}\hat{\lambda}_i$ converges to (a) $\varphi_i(\mathbf{\Phi}) + \tau^2$ for $i = 1, \ldots, m$ and to (b) τ^2 for $i = m+1, \ldots, n-K$ in probability.*
(ii) *$\hat{\mathbf{u}}_i'\mathbf{u}_j$ converges to (a) $v_{ij}(\mathbf{\Phi})\sqrt{\varphi_i(\mathbf{\Phi})/(\varphi_i(\mathbf{\Phi}) + \tau^2)}$ for $i, j = 1, \ldots, m$ and to (b) 0 for $i = m+1, \ldots, n-K$ and $j = 1, \ldots, m$ in probability.*

Proof of Lemma 4 The sample within-scatter matrix $\mathbf{S}_W$ shares its nonzero eigenvalues with the $n \times n$ dual matrix $\mathbf{S}_D = \tilde{\mathbf{X}}'\tilde{\mathbf{X}}$. Under Assumptions 1–4, Theorem 1 of Jung and Marron (2009) gives

$$\mathbf{S}_D/p = (\mathbf{I}_n - \mathbf{J})\mathbf{Z}'\mathbf{\Lambda}\mathbf{Z}(\mathbf{I}_n - \mathbf{J})/p \xrightarrow{P} (\mathbf{I}_n - \mathbf{J})(\mathbf{W}\mathbf{W}' + \tau^2\mathbf{I}_n)(\mathbf{I}_n - \mathbf{J}) =: \mathbf{S}_0 \tag{A3}$$

as $p \to \infty$. Hence, we get $p^{-1}\varphi_i(\mathbf{S}_D) \xrightarrow{P} \varphi_i(\mathbf{S}_0)$ and $v_i(\mathbf{S}_D) \xrightarrow{P} v_i(\mathbf{S}_0)$ for $i = 1, \ldots, n-K$ as $p \to \infty$. Since $\mathbf{\Phi} = \mathbf{W}'(\mathbf{I}_n - \mathbf{J})\mathbf{W}$ is of rank m with probability 1 and $(\mathbf{I}_n - \mathbf{J})\mathbf{W}\mathbf{W}'(\mathbf{I}_n - \mathbf{J})$ shares its nonzero eigenvalues with $\mathbf{\Phi}$, we have (b) of (i). To show (a) of (i), we claim that for $i = 1, \ldots, m$, $\varphi_i(\mathbf{S}_0) = \varphi_i(\mathbf{\Phi}) + \tau^2$. To show this, let λ be a nonzero eigenvalue of $(\mathbf{I}_n - \mathbf{J})\mathbf{W}\mathbf{W}'(\mathbf{I}_n - \mathbf{J})$ and $\mathbf{v}$ be its corresponding eigenvector. Then, there exists $\mathbf{u} \in \mathbb{R}^n$ satisfying $\mathbf{v} = (\mathbf{I}_n - \mathbf{J})\mathbf{u}$, and thus $(\mathbf{I}_n - \mathbf{J})\mathbf{W}\mathbf{W}'(\mathbf{I}_n - \mathbf{J})\mathbf{u} = \lambda(\mathbf{I}_n - \mathbf{J})\mathbf{u}$. Hence,

$$\mathbf{S}_0\mathbf{v} = (\mathbf{I}_n - \mathbf{J})(\mathbf{W}\mathbf{W}' + \tau^2\mathbf{I}_n)(\mathbf{I}_n - \mathbf{J})\mathbf{u} = (\lambda + \tau^2)(\mathbf{I}_n - \mathbf{J})\mathbf{u} = (\lambda + \tau^2)\mathbf{v}$$

and we have $\varphi_i(\mathbf{S}_0) = \varphi_i(\mathbf{\Phi}) + \tau^2$ for $i = 1, \ldots, m$.

Now, we aim to show (b). Note that $v_i(\mathbf{\Phi})$, $v_i(\mathbf{S}_0)$, and $\sqrt{\varphi_i(\mathbf{\Phi})}$ are the ith right-singular vector, left-singular vector, and singular value of $(\mathbf{I}_n - \mathbf{J})\mathbf{W}$, respectively. Therefore, we have $(\mathbf{I}_n - \mathbf{J})\mathbf{W}v_i(\mathbf{\Phi}) = \varphi_i(\mathbf{\Phi})^{1/2}v_i(\mathbf{S}_0)$ for $i = 1, \ldots, n-K$. Hence, for $i, j = 1, \ldots, m$,

$$\begin{aligned}\mathbf{u}_j'\hat{\mathbf{u}}_i = \hat{\lambda}_i^{-1/2}\lambda_j^{1/2}\mathbf{z}_j'(\mathbf{I}_n - \mathbf{J})v_i(\mathbf{S}_D) &\xrightarrow{P} \frac{\sigma_j\mathbf{z}_j'(\mathbf{I}_n - \mathbf{J})}{\sqrt{\varphi_i(\mathbf{\Phi}) + \tau^2}}\frac{(\mathbf{I}_n - \mathbf{J})\mathbf{W}}{\sqrt{\varphi_i(\mathbf{\Phi})}}v_i(\mathbf{\Phi}) \\ &= \frac{\mathbf{e}_{m,j}'\mathbf{\Phi}v_i(\mathbf{\Phi})}{\sqrt{\varphi_i(\mathbf{\Phi})(\varphi_i(\mathbf{\Phi}) + \tau^2)}} = \frac{\varphi_i(\mathbf{\Phi})\mathbf{e}_{m,j}'v_i(\mathbf{\Phi})}{\sqrt{\varphi_i(\mathbf{\Phi})(\varphi_i(\mathbf{\Phi}) + \tau^2)}} = \sqrt{\frac{\varphi_i(\mathbf{\Phi})}{\varphi(\mathbf{\Phi}) + \tau^2}}v_{ij}(\mathbf{\Phi})\end{aligned}$$

as $p \to \infty$. Here, $\mathbf{e}_{m,j}$ is the jth standard vector in $\mathbb{R}^m$. Meanwhile, since $(\mathbf{I}_n - \mathbf{J})\mathbf{W}\mathbf{W}'(\mathbf{I}_n - \mathbf{J})$ is of rank m with probability 1, $v_i(\mathbf{S}_0)$ is orthogonal to the column space of $(\mathbf{I}_n - \mathbf{J})\mathbf{W}\mathbf{W}'(\mathbf{I}_n - \mathbf{J})$ for $i = m+1, \ldots, n-K$. In other words, $v_i(\mathbf{S}_0)'(\mathbf{I}_n - \mathbf{J})\mathbf{W} = 0$ for $i = m+1, \ldots, n-K$. Hence, for $i = m+1, \ldots, n-K$ and $j = 1, \ldots, m$,

$$\mathbf{u}_j'\hat{\mathbf{u}}_i \xrightarrow{P} \frac{\sigma_j\mathbf{z}_j'(\mathbf{I}_n - \mathbf{J})}{\sqrt{\varphi_i(\mathbf{\Phi})}}v_i(\mathbf{S}_0) = \frac{\mathbf{e}_{m,j}'\mathbf{W}'(\mathbf{I}_n - \mathbf{J})}{\sqrt{\varphi_i(\mathbf{\Phi})}}v_i(\mathbf{S}_0) = 0$$

as $p \to \infty$.

Throughout, we assume $\boldsymbol{\mu}_K = \mathbf{0}_p$ without loss of generality. For $1 \le i \le K-1$, there exists $\delta_i > 0$ such that $p^{-1/2}\|\boldsymbol{\mu}_i\| \to \delta_i$ as $p \to \infty$. Also, for $1 \le i, j \le K-1$, there exists $\eta_{i,j} \in \mathbb{R}$ such that $p^{-1}\boldsymbol{\mu}_i'\boldsymbol{\mu}_j \to \eta_{i,j}$ as $p \to \infty$. For each $1 \le i \le m$ and $1 \le j \le K-1$, there exists $\theta_{i,j} \in [0, \pi]$ such that $\text{Angle}(\mathbf{u}_i, \boldsymbol{\mu}_j) \to \theta_{i,j}$ as $p \to \infty$. Recall that we denote the sample mean difference matrix as $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{K-1}]$ where $\mathbf{d}_j = \bar{X}_j - \bar{X}_K$ for $j = 1, \ldots, K-1$. Then note that $\mathcal{W}_{MDP} = \text{span}\{(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D}\}$.

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

(i) *For* $j = 1, \ldots, K-1$,

$$p^{-1/2}\boldsymbol{\mu}_j'\hat{\mathbf{u}}_i \xrightarrow{P} \sqrt{\frac{\varphi_i(\boldsymbol{\Phi})}{\varphi_i(\boldsymbol{\Phi})+\tau^2}} \sum_{l=1}^{m} \cos\theta_{l,j}\delta_j v_{il}(\boldsymbol{\Phi})$$

if $i = 1, \ldots, m$ and $p^{-1/2}\boldsymbol{\mu}_j'\hat{\mathbf{u}}_i \xrightarrow{P} 0$ if $i = m+1, \ldots, n-K$.

(ii) *For* $j = 1, \ldots, K-1$,

$$p^{-1/2}\mathbf{d}_j'\hat{\mathbf{u}}_i \xrightarrow{P} \sqrt{\frac{\varphi_i(\boldsymbol{\Phi})}{\varphi_i(\boldsymbol{\Phi})+\tau^2}} \sum_{l=1}^{m} \{\sigma_l(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,j}\delta_j\} v_{il}(\boldsymbol{\Phi})$$

if $i = 1, \ldots, m$ and $p^{-1/2}\mathbf{d}_j'\hat{\mathbf{u}}_i \xrightarrow{P} 0$ if $i = m+1, \ldots, n-K$.

(iii) $\mathbf{D}'(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D}/p \xrightarrow{P} \boldsymbol{\Omega}$ *where* $\boldsymbol{\Omega} = (\omega_{i,j}) \in \mathbb{R}^{(K-1)\times(K-1)}$ *and*

$$\omega_{i,j} = k_{i,j} + (n_i^{-1}I_{i,j} + n_K^{-1})\tau^2 + \left[\sum_{l=1}^{m} \frac{\tau^2}{\varphi_l(\boldsymbol{\Phi})+\tau^2} \times \sum_{l=1}^{m}\sum_{l'=1}^{m} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\}\{\sigma_{l'}(\bar{\mathbf{z}}_{j,l'} - \bar{\mathbf{z}}_{K,l'}) + \cos\theta_{l',j}\delta_j\} v_{il}(\boldsymbol{\Phi})v_{il'}(\boldsymbol{\Phi})\right]$$

for $i, j = 1, \ldots, K-1$. Here, $k_{i,j} = \lim_{p\to\infty} p^{-1}\boldsymbol{\mu}_i' P_{\mathcal{U}_m^\perp}\boldsymbol{\mu}_j = \eta_{i,j} - \sum_{l=1}^{m} \cos\theta_{l,i}\cos\theta_{l,j}\delta_i\delta_j$, and $I_{i,j} = 1$ if $i = j$ and $I_{i,j} = 0$ if $i \neq j$.

Proof of Lemma 5 First, we extend the results of Lemma C.1 (b) of Chang et al. (2021) which leads that

$$p^{-1}\boldsymbol{\mu}_j'\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{Z} \xrightarrow{P} \sum_{l=1}^{m} \sigma_l \cos\theta_{l,j}\delta_j\mathbf{z}_l'$$

as $p \to \infty$. Combining this with Lemma 4 gives that for $j = 1, \ldots, K-1$,

$$\begin{aligned} p^{-1/2}\boldsymbol{\mu}_j'\hat{\mathbf{u}}_i &= p^{-1/2}\hat{\lambda}_i^{-1/2}\boldsymbol{\mu}_j'\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{Z}(\mathbf{I}_n - \mathbf{J})v_i(\mathbf{S}_D) \\ &\xrightarrow{P} \sum_{l=1}^{m} \cos\theta_{l,j}\delta_j \frac{\sigma_l\mathbf{z}_l'(\mathbf{I}_n - \mathbf{J})}{\sqrt{\varphi_i(\boldsymbol{\Phi})+\tau^2}} \frac{(\mathbf{I}_n - \mathbf{J})\mathbf{W}}{\sqrt{\varphi_i(\boldsymbol{\Phi})}} v_i(\boldsymbol{\Phi}) \\ &= \sqrt{\frac{\varphi_i(\boldsymbol{\Phi})}{\varphi_i(\boldsymbol{\Phi})+\tau^2}} \sum_{l=1}^{m} \cos\theta_{l,j}\delta_j v_{il}(\boldsymbol{\Phi}) \end{aligned} \tag{A4}$$

as $p \to \infty$ and we have (i). Next, let $\mathbf{a}_j = (a_{1j}, \ldots, a_{nj})'$ where

$$a_{ij} = \begin{cases} 1/n_j, & \text{if } i \in (\sum_{k=1}^{j-1} n_k, \sum_{k=1}^{j} n_k] \\ -1/n_K, & \text{if } i \in (\sum_{k=1}^{K-1} n_k, n] \\ 0, & o.w \end{cases}$$

for $j = 1, \ldots, K-1$ so that $\mathbf{d}_j = \mathbf{X}\mathbf{a}_j = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{a}_j + \boldsymbol{\mu}_j$. It can be shown using Lemma C.1 (d) of Chang et al. (2021) that

$$\begin{aligned} p^{-1/2}\mathbf{a}_j'\mathbf{Z}'\mathbf{\Lambda}^{1/2}\mathbf{U}'\hat{\mathbf{u}}_i &= p^{-1/2}\hat{\lambda}_i^{-1/2}\mathbf{a}_j'\mathbf{Z}'\mathbf{\Lambda}\mathbf{Z}(\mathbf{I}_n - \mathbf{J})v_i(\mathbf{S}_D) \\ &\xrightarrow{P} \sum_{l=1}^{m} \sigma_l \mathbf{a}_j'\mathbf{z}_l \frac{\sigma_l \mathbf{z}_l'(\mathbf{I}_n - \mathbf{J})}{\sqrt{\varphi_i(\mathbf{\Phi}) + \tau^2}} \frac{(\mathbf{I}_n - \mathbf{J})\mathbf{W}}{\sqrt{\varphi_i(\mathbf{\Phi})}} v_i(\mathbf{\Phi}) \\ &= \sqrt{\frac{\varphi_i(\mathbf{\Phi})}{\varphi_i(\mathbf{\Phi}) + \tau^2}} \sum_{l=1}^{m} \sigma_l(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l})v_{il}(\mathbf{\Phi}) \end{aligned} \tag{A5}$$

for $j = 1, \ldots, K-1$ as $p \to \infty$. Adding (A4) and (A5) gives that

$$p^{-1/2}\mathbf{d}_j'\hat{\mathbf{u}}_i \xrightarrow{P} \sqrt{\frac{\varphi_i(\mathbf{\Phi})}{\varphi_i(\mathbf{\Phi}) + \tau^2}} \sum_{l=1}^{m} \{\sigma_l(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,j}\delta_j\} v_{il}(\mathbf{\Phi}) \tag{A6}$$

as $p \to \infty$ and we have (ii).

Next, note that $\mathbf{D}'(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D}/p$ is a $(K-1)\times(K-1)$ matrix whose (i, j)-coordinate is $p^{-1}\mathbf{d}_i'(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{d}_j$ for $i, j = 1, \ldots, K-1$. Lemma C.1 (b) and (d) of Chang et al. (2021) give

$$\begin{aligned} & p^{-1}\mathbf{d}_i'\mathbf{d}_j \\ &= p^{-1}\left(\boldsymbol{\mu}_i'\boldsymbol{\mu}_j + \boldsymbol{\mu}_i'\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{a}_j + \boldsymbol{\mu}_j'\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Z}\mathbf{a}_i + \mathbf{a}_i'\mathbf{Z}'\mathbf{\Lambda}\mathbf{Z}\mathbf{a}_j\right) \\ &\xrightarrow{P} \eta_{i,j} + \sum_{l=1}^{m} \sigma_l(\cos\theta_{l,i}\delta_i\mathbf{z}_l'\mathbf{a}_j + \cos\theta_{l,j}\delta_j\mathbf{z}_l'\mathbf{a}_i + \sigma_l\mathbf{a}_i'\mathbf{z}_l\mathbf{z}_l'\mathbf{a}_j) + \tau^2\mathbf{a}_i'\mathbf{a}_j \\ &= \eta_{i,j} + (n_i^{-1}I_{i,j} + n_K^{-1})\tau^2 \\ &\quad + \sum_{l=1}^{m} \sigma_l\{\cos\theta_{l,i}\delta_i(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,j}\delta_j(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l})(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l})\} \end{aligned} \tag{A7}$$

where $I_{i,j} = 1$ if $i = j$ and $I_{i,j} = 0$ if $i \neq j$. Then combining (A7) and (ii) gives $\mathbf{D}'(\mathbf{I} - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D}/p \xrightarrow{P} \mathbf{\Omega}$ where $\mathbf{\Omega} = (\omega_{i,j})$ and

$$
\begin{aligned}
\omega_{i,j} &= \eta_{i,j} + (n_i^{-1} I_{i,j} + n_K^{-1})\tau^2 \\
&+ \sum_{l=1}^{m} \sigma_l \{\cos\theta_{l,i}\delta_i(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,j}\delta_j(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l})(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_{K,l})\} \\
&- \sum_{\iota=1}^{m}\sum_{l=1}^{m}\sum_{l'=1}^{m} \frac{\varphi_\iota(\mathbf{\Phi}) v_{\iota l}(\mathbf{\Phi}) v_{\iota l'}(\mathbf{\Phi})}{\varphi_\iota(\mathbf{\Phi}) + \tau^2} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\}\{\sigma_{l'}(\bar{\mathbf{z}}_{j,l'} - \bar{\mathbf{z}}_{K,l'}) + \cos\theta_{l',j}\delta_j\} \\
&= k_{i,j} + (n_i^{-1} I_{i,j} + n_K^{-1})\tau^2 + \left[\sum_{\iota=1}^{m} \frac{\tau^2}{\varphi_\iota(\mathbf{\Phi}) + \tau^2}\right. \\
&\left. \times \sum_{l=1}^{m}\sum_{l'=1}^{m} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\}\{\sigma_{l'}(\bar{\mathbf{z}}_{j,l'} - \bar{\mathbf{z}}_{K,l'}) + \cos\theta_{l',j}\delta_j\} v_{\iota l}(\mathbf{\Phi}) v_{\iota l'}(\mathbf{\Phi})\right]
\end{aligned}
$$

for $i,j = 1,\ldots,K-1$ as $p \to \infty$ where $k_{i,j} = \eta_{i,j} - \sum_{l=1}^{m} \cos\theta_{l,i}\cos\theta_{l,j}\delta_i\delta_j$. The last equality is due to the fact that $\sum_{\iota=1}^{m} v_{\iota l}(\mathbf{\Phi}) v_{\iota l'}(\mathbf{\Phi}) = I_{l,l'}$.

Write the singular value decomposition $(\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W')\mathbf{D} = \sum_{i=1}^{K-1} \theta_i \mathbf{p}_i \mathbf{q}_i'$ where θ_i is the ith largest nonzero singular value and $\mathbf{p}_i$ (or $\mathbf{q}_i$) is the corresponding left (or right) singular vector. Then, for $i = 1,\ldots,K-1$,

$$
\theta_i^2/p \xrightarrow{P} \varphi_i(\mathbf{\Omega}), \tag{A8}
$$

and

$$
\mathbf{q}_i \xrightarrow{P} v_i(\mathbf{\Omega}), \tag{A9}
$$

as $p \to \infty$. These convergences play a key role when analyzing the MDP subspace $\mathcal{W}_{MDP} = \text{span}(\{\mathbf{p}_i\}_{i=1}^{K-1})$ in the proofs of main results.

Equivalent definitions of second data piling

Recall that the collection of $(K-1)$-dimensional subspaces inducing the second data piling is defined as

$$
\begin{aligned}
\mathfrak{C}_{SDP} = \{\mathcal{V} \subset \mathcal{S}_X :\ &\dim(\mathcal{V}) = K-1 \text{ and for any } Y, Y' \in \mathcal{Y} \text{ with } \pi(Y) = \pi(Y'), \\
&p^{-1/2}\left\|P_{\mathcal{V}}(Y - Y')\right\|_2 \xrightarrow{P} 0 \text{ as } p \to \infty\}.
\end{aligned}
$$

In this section, we introduce equivalent definitions of $\mathfrak{C}_{SDP}$ to provide a clear understanding of second data piling subspaces.

Lemma 6 *Suppose that Assumptions 1–4 hold. Let*

$$\mathfrak{C}_{SDP,2} = \{\mathcal{V} \subset \mathcal{S}_X : \dim(\mathcal{V}) = K-1 \text{ and for any } Y, Y' \in \mathcal{Y} \text{ with } \pi(Y) = \pi(Y'), \\ p^{-1/2}\|P_{\mathcal{V}}(Y-Y')\|_2 \xrightarrow{L^2} 0 \text{ as } p \to \infty\},$$

$$\mathfrak{C}_{SDP,3} = \left\{\mathcal{V} \subset \mathcal{S}_X : \dim(\mathcal{V}) = K-1 \text{ and } \|P_{\mathcal{V}}\mathbf{u}_i\|_2 \xrightarrow{P} 0 \text{ as } p \to \infty \text{ for } i = 1, \ldots, m\right\},$$

$$\mathfrak{C}_{SDP,4} = \left\{\mathcal{V} \subset \mathcal{S}_X : \dim(\mathcal{V}) = K-1 \text{ and } \|P_{\mathcal{V}}\mathbf{u}_i\|_2 \xrightarrow{L^2} 0 \text{ as } p \to \infty \text{ for } i = 1, \ldots, m\right\}.$$

Then, $\mathfrak{C}_{SDP} = \mathfrak{C}_{SDP,2} = \mathfrak{C}_{SDP,3} = \mathfrak{C}_{SDP,4}$.

Proof of Lemma 6 First, Chebyshev's inequality gives $\mathfrak{C}_{SDP,2} \subset \mathfrak{C}_{SDP}$ and $\mathfrak{C}_{SDP,4} \subset \mathfrak{C}_{SDP,3}$. To show $\mathfrak{C}_{SDP} \subset \mathfrak{C}_{SDP,2}$, we claim that for any $Y, Y' \in \mathcal{Y}$ with $\pi(Y) = \pi(Y')$, a sequence $\{p^{-1/2}\|P_{\mathcal{V}}(Y-Y')\|_2\}$ is uniformly integrable. We have

$$p^{-1}\mathbb{E}\left(\|P_{\mathcal{V}}(Y-Y')\|_2^2\right) \le p^{-1}\mathbb{E}\left(\|Y-Y'\|_2^2\right) = p^{-1}\,\text{trace}\,(2\mathbf{\Sigma}) \le 2\left(\sum_{i=1}^{m}\sigma_i^2 + M\right).$$

where $M > 0$ is an upper bound of $\{\lambda_i\}_{i=m+1}^{p}$. Since a sequence with uniformly bounded second moments is uniformly integrable, from Vitali's convergence theorem (Bogachev, 2007), we have $\mathfrak{C}_{SDP} \subset \mathfrak{C}_{SDP,2}$. Similarly, $\mathbb{E}(\|P_{\mathcal{V}}\mathbf{u}_i\|_2^2) \le 1$ guarantees that $\mathfrak{C}_{SDP,3} \subset \mathfrak{C}_{SDP,4}$. Hence, $\mathfrak{C}_{SDP} = \mathfrak{C}_{SDP,2}$ and $\mathfrak{C}_{SDP,3} = \mathfrak{C}_{SDP,4}$.

Next, we aim to show that $\mathfrak{C}_{SDP,2} = \mathfrak{C}_{SDP,4}$. For any $\mathcal{V} \in \mathfrak{C}_{SDP,2}$ and $1 \le i \le m$, we have $p^{-1/2}\left\|P_{\mathcal{V}}\mathbf{u}_i'(Y-Y')\mathbf{u}_i\right\|_2 \xrightarrow{L^2} 0$ as $p \to \infty$ for any $Y, Y' \in \mathcal{Y}$ with $\pi(Y) = \pi(Y')$. Then Lemma 3.3 of Chang et al. (2021) gives that $\mathbf{u}_i'P_{\mathcal{V}}\mathbf{u}_i \xrightarrow{L^2} 0$ as $p \to \infty$, hence, $\mathfrak{C}_{SDP,2} \subset \mathfrak{C}_{SDP,4}$. To show $\mathfrak{C}_{SDP,4} \subset \mathfrak{C}_{SDP,2}$, write the principal component decomposition of $Y - Y'$ as $\sum_{i=1}^{p}\lambda_i^{1/2}(\zeta_i - \zeta_i')\mathbf{u}_i$ where ζ_i and ζ_i' denote the ith principal component of Y and Y', respectively. Then, ζ_i, ζ_i' are independent to the training data and to each other, and have mean zero and unit variance. Also, we can write

$$\begin{aligned} p^{-1/2}P_{\mathcal{V}}(Y-Y') &= \sum_{i=1}^{m}\sigma_i P_{\mathcal{V}}\mathbf{u}_i(\zeta_i - \zeta_i') + p^{-1/2}\sum_{i=m+1}^{p}\lambda_i^{1/2}P_{\mathcal{V}}\mathbf{u}_i(\zeta_i - \zeta_i') \\ &=: A_1 + A_2. \end{aligned}$$

Here, it suffices to show that $\|A_2\|_2 \xrightarrow{L^2} 0$ as $p \to \infty$, which is followed by

$$\begin{aligned} &\mathbb{E}(\|A_2\|_2^2) = p^{-1}\mathbb{E}\left\|\sum_{i=m+1}^{p}\lambda_i^{1/2}P_{\mathcal{V}}\mathbf{u}_i(\zeta_i - \zeta_i')\right\|_2^2 \\ &\le p^{-1}\mathbb{E}\left\{\sum_{i=m+1}^{p}\lambda_i\|P_{\mathcal{V}}\mathbf{u}_i\|_2^2(\zeta_i - \zeta_i')^2 + \sum_{i\neq t}\lambda_i^{1/2}\lambda_t^{1/2}\|P_{\mathcal{V}}\mathbf{u}_i\|_2\|P_{\mathcal{V}}\mathbf{u}_t\|_2(\zeta_i - \zeta_i')(\zeta_t - \zeta_t')\right\} \\ &\le p^{-1}2M\mathbb{E}\left(\sum_{i=m+1}^{p}\|P_{\mathcal{V}}\mathbf{u}_i\|_2^2\right) \le p^{-1}2M\mathbb{E}\{\text{ trace }(P_{\mathcal{V}})\} = p^{-1}2(K-1)M \to 0 \end{aligned}$$

as $p \to \infty$.

Proofs of main results

Proof of Lemma 1

Proof Since $\mathbf{S}_W$ has rank at most $n-K$, its orthogonal complement space has rank $q \geq p-(n-K)$. Let $\hat{\mathbf{U}}_{W^\perp}$ be an orthonormal basis of the orthogonal complement space. Then the constraint trace $(\mathbf{V}'\mathbf{S}_W\mathbf{V}) = 0$ requires that candidates of (3) is of the form $\mathbf{V} = \hat{\mathbf{U}}_{W^\perp}\mathbf{O}$ for $\mathbf{O} \in \mathcal{O}(q,d)$. Thus, (3) becomes, writing $\mathbf{S}_{B\setminus W} = \hat{\mathbf{U}}'_{W^\perp}\mathbf{S}_B\hat{\mathbf{U}}_{W^\perp}$,

$$\max_{\mathbf{O}\in\mathcal{O}(q,d)} \text{trace}\,(\mathbf{O}'\mathbf{S}_{B\setminus W}\mathbf{O}) = \sum_{i=1}^{d} \varphi_i(\mathbf{S}_{B\setminus W}),$$

and the maximum is achieved by choosing the columns of $\mathbf{O}$ as the d leading eigenvectors of $\mathbf{S}_{B\setminus W}$ (Corollary 4.3.39, Horn & Johnson, 2012). Therefore

$$\tilde{\mathbf{V}}_d = \hat{\mathbf{U}}_{W^\perp}[v_1(\mathbf{S}_{B\setminus W}), \ldots, v_d(\mathbf{S}_{B\setminus W})]. \tag{A10}$$

Note that $\hat{\mathbf{U}}_{W^\perp}\mathbf{S}_{B\setminus W}\hat{\mathbf{U}}'_{W^\perp} = \hat{\mathbf{U}}_{W^\perp}\hat{\mathbf{U}}'_{W^\perp}\mathbf{M}\mathbf{M}'\hat{\mathbf{U}}_{W^\perp}\hat{\mathbf{U}}'_{W^\perp}$ and that, for $\tilde{\mathbf{V}}_d$ defined in (A10), it can be checked that

$$\tilde{\mathbf{V}}'_d\hat{\mathbf{U}}_{W^\perp}\mathbf{S}_{B\setminus W}\hat{\mathbf{U}}'_{W^\perp}\tilde{\mathbf{V}}_d = \text{diag}\,(\varphi_1(\mathbf{S}_{B\setminus W}), \ldots, \varphi_d(\mathbf{S}_{B\setminus W})).$$

Thus, the columns of $\tilde{\mathbf{V}}_d$ are the d leading eigenvectors of $\hat{\mathbf{U}}_{W^\perp}\mathbf{S}_{B\setminus W}\hat{\mathbf{U}}'_{W^\perp}$ whose eigenvalues are exactly $\varphi_i(\mathbf{S}_{B\setminus W})$'s. Equivalently, $\tilde{\mathbf{V}}_d$ consists of the left singular vectors of $\hat{\mathbf{U}}_{W^\perp}\hat{\mathbf{U}}'_{W^\perp}\mathbf{M}$ corresponding to the d largest singular values. Note that $\hat{\mathbf{U}}_{W^\perp}\hat{\mathbf{U}}'_{W^\perp}$ is the projection matrix onto the orthogonal complement space of $\mathbf{S}_W$. Since $\mathbf{S}_W = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$, the orthogonal complement space of $\mathbf{S}_W$ is the orthogonal complement space of $\tilde{\mathbf{X}}$. Thus, $\hat{\mathbf{U}}_{W^\perp}\hat{\mathbf{U}}'_{W^\perp}\mathbf{M} = (\mathbf{I}_p - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\dagger)\mathbf{M} = (\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}'_W)\mathbf{M}$. □

Proof of Theorem 1

Proof For any independent observation $Y \in \mathcal{Y}$, assume $\pi(Y) = k$ $(1 \leq k \leq K)$. Decompose $p^{-1/2}(Y-\boldsymbol{\mu}_k)$ into two terms:

$$p^{-1/2}(Y-\boldsymbol{\mu}_k) = p^{-1/2}\sum_{i=1}^{m}\mathbf{u}'_i(Y-\boldsymbol{\mu}_k)\mathbf{u}_i + p^{-1/2}\sum_{i=m+1}^{p}\mathbf{u}'_i(Y-\boldsymbol{\mu}_k)\mathbf{u}_i \tag{A11}$$

Projecting (A11) onto $\mathcal{S} = \mathcal{W}_{MDP} \oplus \text{span}(\{\hat{\mathbf{u}}_i\}_{i=1}^m)$, we have

$$p^{-1/2}P_{\mathcal{S}}(Y-\boldsymbol{\mu}_k) = p^{-1/2}\sum_{i=1}^{m}\mathbf{u}'_i(Y-\boldsymbol{\mu}_k)P_{\mathcal{S}}\mathbf{u}_i + p^{-1/2}\sum_{i=m+1}^{p}\mathbf{u}'_i(Y-\boldsymbol{\mu}_k)P_{\mathcal{S}}\mathbf{u}_i.$$

Note that for $i = 1, \ldots, m$, $P_{\mathcal{S}}\mathbf{u}_i = \mathbf{u}_{i,\mathcal{S}} \in \mathcal{T} = \text{span}(\{\mathbf{u}_{j,\mathcal{S}}\}_{j=1}^{m})$. Hence, we claim that the second term in the above equation degenerates as $p \to \infty$. Let $\hat{\mathbf{U}}_1 = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}_m}]$, $\mathbf{U}_1 = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ and $\mathbf{U}_2 = [\mathbf{u}_{m+1}, \ldots, \mathbf{u}_p]$. It suffices to show that

$$p^{-1/2}\hat{\mathbf{U}}_1'\mathbf{U}_2\mathbf{U}_2'(Y - \boldsymbol{\mu}_k) \xrightarrow{P} 0 \tag{A12}$$

and

$$p^{-1/2}\mathbf{p}_i'\mathbf{U}_2\mathbf{U}_2'(Y - \boldsymbol{\mu}_k) \xrightarrow{P} 0 \tag{A13}$$

for $i = 1, \ldots, K-1$ as $p \to \infty$.

To show (A12), let $\boldsymbol{\zeta} \in \mathbb{R}^p$ consists of the standardized true principal component scores of Y, that is, $\boldsymbol{\zeta} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}'(Y - \boldsymbol{\mu}_k) = (\zeta_1, \ldots, \zeta_p)'$. Also, write $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1', \boldsymbol{\zeta}_2')'$ where $\boldsymbol{\zeta}_1 \in \mathbb{R}^m$ and $\boldsymbol{\zeta}_2 \in \mathbb{R}^{p-m}$. Note that the elements of $\boldsymbol{\zeta}$ are uncorrelated and have mean zero and unit variance. Also,

$$p^{-1/2}\hat{\mathbf{U}}_1'\mathbf{U}_2\mathbf{U}_2'(Y - \boldsymbol{\mu}_k) = p^{-1/2}\hat{\mathbf{U}}_1'\mathbf{U}_2\mathbf{U}_2'\mathbf{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\zeta} = p^{-1/2}\hat{\mathbf{U}}_1'\mathbf{U}_2\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2$$

where $\boldsymbol{\Lambda}_2 = \text{diag}\,(\lambda_{m+1}, \ldots, \lambda_p)$. The ith coordinate of $\hat{\mathbf{U}}_1'\mathbf{U}_2\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2$ is $\sum_{j=m+1}^{p}(\hat{\mathbf{u}}_i'\mathbf{u}_j)\lambda_j^{1/2}\zeta_j$ for $i = 1, \ldots, m$. Then Chebyshev's inequality gives

$$\begin{aligned}
\mathbb{P}\left(p^{-1/2}\left|\sum_{j=m+1}^{p}(\hat{\mathbf{u}}_i'\mathbf{u}_j)\lambda_j^{1/2}\zeta_j\right| > \epsilon\right) &\le p^{-1}\epsilon^{-2}\mathbb{E}\left\{\left(\sum_{j=m+1}^{p}(\hat{\mathbf{u}}_i'\mathbf{u}_j)\lambda_j^{1/2}\zeta_j\right)^2\right\} \\
&\le p^{-1}\epsilon^{-2}M\sum_{j=m+1}^{p}\mathbb{E}\left\{(\hat{\mathbf{u}}_i'\mathbf{u}_j)^2\right\} \\
&\le p^{-1}\epsilon^{-2}M\mathbb{E}\left(\left\|\hat{\mathbf{u}}_i\right\|_2^2\right) = p^{-1}\epsilon^{-2}M \to 0
\end{aligned}$$

as $p \to \infty$ where $M > 0$ is an upper bound of $\left\{\lambda_i\right\}_{i=m+1}^{p}$, and (A12) follows.

To show (A13), recall the singular value decomposition $(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D} = \sum_{i=1}^{K-1}\theta_i\mathbf{p}_i\mathbf{q}_i'$. For $i = 1, \ldots, K-1$, $\mathbf{p}_i = \theta_i^{-1}(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D}\mathbf{q}_i$ holds. Therefore,

$$
\begin{aligned}
p^{-1/2}\mathbf{p}_i'\mathbf{U}_2\mathbf{U}_2'(Y-\boldsymbol{\mu}_k) &= p^{-1/2}\theta_i^{-1}\mathbf{q}_i'\mathbf{D}'(\mathbf{I}_p-\hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{U}_2\mathbf{U}_2'\mathbf{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\zeta}\\
&= p^{-1/2}\theta_i^{-1}\mathbf{q}_i'\mathbf{D}'(\mathbf{I}_p-\hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{U}_2\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2\\
&= (p^{-1/2}\theta_i)^{-1}\mathbf{q}_i'(\mathbf{A}_1-\mathbf{A}_2),
\end{aligned}
$$

where $\mathbf{A}_1 = p^{-1}\mathbf{D}'\mathbf{U}_2\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2$ and $\mathbf{A}_2 = p^{-1}\mathbf{D}'\hat{\mathbf{U}}_W\hat{\mathbf{U}}_W'\mathbf{U}_2\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2$. For $i = 1,\dots,K-1$, we already know that $p^{-1/2}\theta_i$ and $\mathbf{q}_i$ converge to $\varphi_i(\boldsymbol{\Omega})$ and $v_i(\boldsymbol{\Omega})$, respectively. Therefore, it remains to show that $\mathbf{A}_1-\mathbf{A}_2 \xrightarrow{P} 0$. Meanwhile, we can show that $\mathbf{A}_2 \to 0$ as $p\to\infty$ immediately from Lemma 5 (ii) and (A12). For $\mathbf{A}_1$, denote the ith column of $\mathbf{U}_2'\mathbf{D}$ as $\mathbf{x}^{(i)} = (x_{m+1}^{(i)},\dots,x_p^{(i)})$. Then the ith coordinate of $\mathbf{A}_1$ is $p^{-1}\mathbf{x}^{(i)\prime}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2$ for $i = 1,\dots,K-1$ and we can show that

$$
\begin{aligned}
&\mathbb{E}\left\{\left(p^{-1}\mathbf{x}^{(i)\prime}\boldsymbol{\Lambda}_2^{1/2}\boldsymbol{\zeta}_2\right)^2\right\} = p^{-2}\mathbb{E}\left\{\left(\sum_{j=m+1}^{p} x_j^{(i)}\lambda_j^{1/2}\zeta_j\right)^2\right\}\\
&\le p^{-2}M\mathbb{E}\left\{\sum_{j=m+1}^{p}(x_j^{(i)}\zeta_j)^2 + \sum_{j\neq j'} x_j^{(i)}\zeta_j x_{j'}^{(i)}\zeta_{j'}\right\}\\
&= p^{-2}M\sum_{j=m+1}^{p}\mathbb{E}\left\{(x_j^{(i)})^2\right\} = p^{-2}M\mathbb{E}\left(\left\|\mathbf{x}^{(i)}\right\|_2^2\right)\\
&\le p^{-2}M\left\{\left\|\mathbf{U}_2'\boldsymbol{\mu}_i\right\|_2^2 + (p-m)\left(\frac{1}{n_i}+\frac{1}{n_K}\right)M\right\}.
\end{aligned}
\tag{A14}
$$

Combining Chebyshev's inequality and (A14) gives $\mathbf{A}_1 \xrightarrow{P} 0$ as $p\to\infty$, and thus (A13) follows. □

Proof of Lemma 2

Proof For any $\mathcal{V}\in\mathfrak{C}_{SDP}$ and $1\le i\le m$,

$$
\begin{aligned}
P_{\mathcal{V}}\mathbf{u}_i &= P_{\mathcal{V}}P_{\mathcal{H}}\mathbf{u}_i + P_{\mathcal{V}}P_{\mathcal{T}}\mathbf{u}_i + P_{\mathcal{V}}P_{\mathcal{S}^\perp}\mathbf{u}_i\\
&= P_{\mathcal{V}}P_{\mathcal{T}}\mathbf{u}_{i,\mathcal{S}} + P_{\mathcal{V}}P_{\mathcal{S}^\perp}\mathbf{u}_i.
\end{aligned}
$$

The second equality is due to $P_{\mathcal{H}}\mathbf{u}_i = P_{\mathcal{T}^\perp|_{\mathcal{S}}}\mathbf{u}_i = P_{\mathcal{T}^\perp|_{\mathcal{S}}}P_{\mathcal{S}}\mathbf{u}_i = P_{\mathcal{T}^\perp|_{\mathcal{S}}}\mathbf{u}_{i,\mathcal{S}} = \mathbf{0}_p$. Also, Lemma 4 (ii) gives $\mathbf{u}_i'\hat{\mathbf{u}}_j \xrightarrow{P} 0$ as $p\to\infty$ for $j = m+1,\dots,n-K$, and thus $\left\|P_{\mathcal{S}^\perp}\mathbf{u}_i\right\|_2\to 0$ as $p\to\infty$. Note that $\left\|P_{\mathcal{V}}\mathbf{u}_i\right\|_2 \xrightarrow{P} 0$ as $p\to\infty$ for $1\le i\le m$ if and only

if $\|P_{\mathcal{V}}\mathbf{u}_{i,\mathcal{S}}\|_2 \xrightarrow{P} 0$ as $p \to \infty$ for $1 \le i \le m$. Now, Lemma 6 gives that $\mathcal{V} \in \mathfrak{C}_{SDP}$ if and only if $\|P_{\mathcal{V}}\mathbf{u}_{i,\mathcal{S}}\|_2 \xrightarrow{P} 0$ as $p \to \infty$.

Next, for any given $\mathcal{V} \in \mathfrak{C}_{SDP}$, let $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{K-1}\}$ be an orthonormal basis for $\mathcal{V}$. Write $\mathcal{B}_p = \mathcal{H} \oplus \mathcal{S}^{\perp}$. Then Lemma 3.3 of Chang et al. (2021) and $\left\|P_{\mathcal{V}}\mathbf{u}_{j,\mathcal{S}}\right\|_2 \xrightarrow{P} 0$ for all $j = 1, \ldots, m$ lead to $\left\|P_{\mathcal{B}_p}\mathbf{v}_i\right\|_2 \xrightarrow{P} 1$ and $\left\|(\mathbf{I}_p - P_{\mathcal{B}_p})\mathbf{v}_i\right\|_2 \xrightarrow{P} 0$ as $p \to \infty$ for $i = 1, \ldots, K-1$. Moreover, for $1 \le i \ne j \le K-1$, $\mathbf{v}_i' P_{\mathcal{B}_p}\mathbf{v}_j = -\mathbf{v}_i'(\mathbf{I}_p - P_{\mathcal{B}_p})\mathbf{v}_j \xrightarrow{P} 0$ as $p \to \infty$. Let $\tilde{\mathcal{V}} = \text{span}(\{P_{\mathcal{B}_p}\mathbf{v}_i\}_{i=1}^{K-1}) \subset \mathcal{B}_p$ and $\tilde{\mathbf{v}}_i = \left\|(\mathbf{I}_p - \sum_{k=1}^{i-1} P_{\tilde{\mathbf{v}}_k})P_{\mathcal{B}_p}\mathbf{v}_i\right\|_2^{-1}(\mathbf{I}_p - \sum_{k=1}^{i-1} P_{\tilde{\mathbf{v}}_k})P_{\mathcal{B}_p}\mathbf{v}_i$ so that $\tilde{\mathbf{V}} = [\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_{K-1}]$ forms an orthonormal basis for $\tilde{\mathcal{V}}$. Then $\mathbf{V}'\tilde{\mathbf{V}} \xrightarrow{P} \mathbf{I}_{K-1}$ and $\text{Angle}(\mathcal{V}, \tilde{\mathcal{V}}) \xrightarrow{P} 0$ as $p \to \infty$. □

Proof of Theorem 2

Proof For any $i \ne j$ and $\mathcal{V} \in \mathfrak{C}_{SDP}$ such that $D_{ij}(\mathcal{V})$ exists, the triangle inequality gives

$$\begin{aligned} & p^{-1/2}\left|\left\|P_{\mathcal{V}}\boldsymbol{\mu}_{ij}\right\|_2 - \left\|P_{\mathcal{V}}(Y_i - Y_j)\right\|_2\right| \\ & \le p^{-1/2}\left\|P_{\mathcal{V}}(Y_i - \boldsymbol{\mu}_i)\right\|_2 + p^{-1/2}\left\|P_{\mathcal{V}}(Y_j - \boldsymbol{\mu}_j)\right\|_2 \end{aligned}$$

where $Y_i, Y_j \in \mathcal{Y}$ with $\pi(Y_i) = i$, $\pi(Y_j) = j$ and $\boldsymbol{\mu}_{ij} := \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$. It implies that $p^{-1/2}\left\|P_{\mathcal{V}}\boldsymbol{\mu}_{ij}\right\|_2 \to D_{ij}(\mathcal{V})$ as $p \to \infty$. Recall that we decompose the sample space $\mathcal{S}_X$ into three subspaces $\mathcal{T}$, $\mathcal{H}$ and $\mathcal{S}^{\perp}$. Denote $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \mathcal{V}_3$ where $\mathcal{V}_1 = \mathcal{V} \cap \mathcal{T}$, $\mathcal{V}_2 = \mathcal{V} \cap \mathcal{H}$, and $\mathcal{V}_3 = \mathcal{V} \cap \mathcal{S}^{\perp}$. Since $P_{\mathcal{V}} = P_{\mathcal{V}_1} \oplus P_{\mathcal{V}_2} \oplus P_{\mathcal{V}_3}$, the triangle inequality gives that

$$p^{-1/2}\left\|P_{\mathcal{V}}(Y_i - Y_j)\right\|_2 \le \sum_{k=1}^{3} p^{-1/2}\left\|P_{\mathcal{V}_k}(Y_i - Y_j)\right\|_2.$$

Lemma 2 gives that $p^{-1/2}\left\|P_{\mathcal{V}_1}(Y_i - Y_j)\right\|_2 \xrightarrow{P} 0$ as $p \to \infty$. Also, the fact that $\mathcal{V}_2 \le \mathcal{H}$ leads that $p^{-1/2}\left\|P_{\mathcal{V}_2}(Y_i - Y_j)\right\|_2 \le p^{-1/2}\left\|P_{\mathcal{H}}(Y_i - Y_j)\right\|_2$. We now claim that $p^{-1/2}\left\|P_{\mathcal{V}_3}(Y_i - Y_j)\right\|_2 \xrightarrow{P} 0$ as $p \to \infty$. Recall that $\mathcal{V}_3 \le \mathcal{S}^{\perp}$ and the columns of $\hat{\mathbf{U}}_2 = [\hat{\mathbf{u}}_{m+1}, \ldots, \hat{\mathbf{u}}_{n-K}]$ forms an orthonormal basis of $\mathcal{S}^{\perp}$. Also, note that $p^{-1/2}\left\|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2'(Y - \mathbb{E}(Y))\right\|_2 \xrightarrow{P} 0$ as $p \to \infty$ for any $Y \in \mathcal{Y}$. Lemma 5 (i) gives that both of $p^{-1/2}\left\|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2'\boldsymbol{\mu}_i\right\|_2$ and $p^{-1/2}\left\|\hat{\mathbf{U}}_2\hat{\mathbf{U}}_2'\boldsymbol{\mu}_j\right\|_2$ converge to 0 in probability. Hence, $p^{-1/2}\left\|P_{\mathcal{V}_3}(Y_i - Y_j)\right\|_2 \le p^{-1/2}\left\|P_{\mathcal{S}^{\perp}}(Y_i - Y_j)\right\|_2 \xrightarrow{P} 0$ as $p \to \infty$. □

Proof of Theorem 3

Proof Let

$$\boldsymbol{\omega}_{\alpha,i} := \frac{1}{\sqrt{p}} P_{\mathcal{S}} \mathbf{L}_{\alpha,i} = \sum_{k=1}^{m} \frac{\alpha_p}{\hat{\lambda}_k + \alpha_p} \left(\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_k{}' \mathbf{d}_i \right) \hat{\mathbf{u}}_k + \frac{1}{\sqrt{p}} (\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W') \mathbf{d}_i$$

for $i = 1, \ldots, K-1$, which is a scaled projection of $\mathbf{L}_{\alpha,i}$ onto $\mathcal{S}$. Note that $\mathcal{H}_\alpha = \text{span}(\{\boldsymbol{\omega}_{\alpha,i}\}_{i=1}^{K-1})$ and $\mathcal{T} = \text{span}(\{\mathbf{u}_{j,\mathcal{S}}\}_{j=1}^{m})$. We claim that for a consistent estimate $\hat{\alpha}$ of $-\tau^2$ and for all $i = 1, \ldots, K-1$ and $j = 1, \ldots, m$,

$$\text{Angle}(\boldsymbol{\omega}_{\hat{\alpha},i}, \mathbf{u}_{j,\mathcal{S}}) = \arccos \left(\frac{\boldsymbol{\omega}_{\hat{\alpha},i}' \mathbf{u}_{j,\mathcal{S}}}{\|\boldsymbol{\omega}_{\hat{\alpha},i}\|_2 \left\|\mathbf{u}_{j,\mathcal{S}}\right\|_2} \right) \xrightarrow{P} \pi/2 \tag{A15}$$

as $p \to \infty$. Then $\text{Angle}(\mathcal{H}_{\hat{\alpha}}, \mathcal{T}) \xrightarrow{P} \pi/2$ and $\mathcal{H}_{\hat{\alpha}} \in \mathfrak{C}_{SDP}$. To show (A15), we will show that $\mathbf{u}_{j,\mathcal{S}}' \boldsymbol{\omega}_{\hat{\alpha},i} = \mathbf{u}_j' \boldsymbol{\omega}_{\hat{\alpha},i} \xrightarrow{P} 0$ for all $j = 1, \ldots, m$, and $\|\boldsymbol{\omega}_{\hat{\alpha},i}\|_2$ and $\left\|\mathbf{u}_{j,\mathcal{S}}\right\|_2$ do not degenerate as $p \to \infty$. First, combining Lemma 4 and Lemma 5 (ii) gives

$$\begin{aligned}
\mathbf{u}_j' \boldsymbol{\omega}_{\hat{\alpha},i} &= \sum_{k=1}^{m} \frac{\hat{\alpha}}{\hat{\lambda}_k/p + \hat{\alpha}} \left(\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_k{}' \mathbf{d}_i \right) (\mathbf{u}_j' \hat{\mathbf{u}}_k) + \frac{1}{\sqrt{p}} \mathbf{u}_j' (\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W') \mathbf{d}_i \\
&= \sum_{k=1}^{m} \frac{\hat{\alpha}}{\hat{\lambda}_k/p + \hat{\alpha}} \left(\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_k{}' \mathbf{d}_i \right) (\mathbf{u}_j' \hat{\mathbf{u}}_k) + \frac{1}{\sqrt{p}} \mathbf{u}_j' \mathbf{d}_i - \sum_{k=1}^{m} \left(\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_k{}' \mathbf{d}_i \right) (\mathbf{u}_j' \hat{\mathbf{u}}_k) + o_p(1) \\
&\xrightarrow{P} \sum_{k=1}^{m} \sum_{l=1}^{m} \frac{-\tau^2}{\varphi_k(\boldsymbol{\Phi}) + \tau^2} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\} v_{kl}(\boldsymbol{\Phi}) v_{kj}(\boldsymbol{\Phi}) + \sigma_j(\bar{\mathbf{z}}_{i,j} - \bar{\mathbf{z}}_{K,j}) \\
&\quad + \cos\theta_{j,i}\delta_i - \sum_{k=1}^{m} \sum_{l=1}^{m} \frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi}) + \tau^2} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\} v_{kl}(\boldsymbol{\Phi}) v_{kj}(\boldsymbol{\Phi}) \\
&= 0
\end{aligned}$$

as $p \to \infty$ for all $j = 1, \ldots, m$. The last equality is due to the fact that $\sum_{k=1}^{m} v_{kl}(\boldsymbol{\Phi}) v_{kj}(\boldsymbol{\Phi}) = I_{l,j}$. Next, combining Lemma 4, Lemma 5, (A8) and (A9) gives

$$\begin{aligned}
\|\boldsymbol{\omega}_{\hat{\alpha},i}\|_2^2 &= \sum_{k=1}^{m} \left(\frac{\hat{\alpha}}{\hat{\lambda}_k/p + \hat{\alpha}} \right)^2 \left(\frac{1}{\sqrt{p}} \hat{\mathbf{u}}_k' \mathbf{d}_i \right)^2 + \frac{1}{p} \mathbf{d}_i' (\mathbf{I}_p - \hat{\mathbf{U}}_W \hat{\mathbf{U}}_W') \mathbf{d}_i \\
&\xrightarrow{P} \sum_{k=1}^{m} \frac{\tau^4}{\varphi_k(\boldsymbol{\Phi})(\varphi_k(\boldsymbol{\Phi}) + \tau^2)} \left[\sum_{l=1}^{m} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\} v_{kl}(\boldsymbol{\Phi}) \right]^2 + \left(\frac{1}{n_i} + \frac{1}{n_K} \right) \tau^2 \\
&\quad + \left(1 - \sum_{l=1}^{m} \cos^2\theta_{l,i} \right) \delta_i^2 + \sum_{k=1}^{m} \frac{\tau^2}{\varphi_k(\boldsymbol{\Phi}) + \tau^2} \left[\sum_{l=1}^{m} \{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\} v_{kl}(\boldsymbol{\Phi}) \right]^2 > 0
\end{aligned}$$

and

$$
\begin{aligned}
\left\|\mathbf{u}_{j,\mathcal{S}}\right\|_2^2 &= \sum_{i=1}^{K-1}(\mathbf{u}_j'\mathbf{p}_i)^2 + \sum_{k=1}^{m}(\mathbf{u}_j'\hat{\mathbf{u}}_k)^2 \\
&= \sum_{i=1}^{K-1}(p^{-1}\theta_i^2)^{-1}\left\{p^{-1/2}\mathbf{u}_j'(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')\mathbf{D}\mathbf{q}_i\right\}^2 + \sum_{k=1}^{m}(\mathbf{u}_j'\hat{\mathbf{u}}_k)^2 \\
&\xrightarrow{P} \sum_{i=1}^{K-1}\frac{1}{\varphi_i(\boldsymbol{\Omega})}\left[\sum_{k=1}^{m}\sum_{l=1}^{m}\sum_{l'=1}^{K-1}\frac{\tau^2}{\varphi_k(\boldsymbol{\Phi})+\tau^2}\{\sigma_l(\bar{\mathbf{z}}_{l',l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,l'}\delta_{l'}\}v_{kl}(\boldsymbol{\Phi})v_{kj}(\boldsymbol{\Phi})v_{il'}(\boldsymbol{\Omega})\right]^2 \\
&\quad + \sum_{k=1}^{m}\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi})+\tau^2}v_{kj}^2(\boldsymbol{\Phi}) > 0
\end{aligned}
$$

as $p \to \infty$ and (A15) follows. Since the $(K-1)$-dimensional subspace $\mathcal{H}_{\hat{\alpha}}$ is included in the $(m+K-1)$-dimensional subspace $\mathcal{S} = \mathcal{T} \oplus \mathcal{H}$ and is asymptotically orthogonal to any direction in the m-dimensional subspace $\mathcal{T}$, one can check that $\text{Angle}(\mathcal{H}_{\hat{\alpha}}, \mathcal{H}) \xrightarrow{P} 0$ as $p \to \infty$. □

Proof of Lemma 3

Proof For any independent observation $Y \in \mathcal{Y}$, assume $\pi(Y) = y$ for some $1 \le y \le K$. To show Lemma 3, we claim the following hold as $p \to \infty$:

(i) For $i = 1, \ldots, K-1$,

$$\frac{1}{\sqrt{p}}\boldsymbol{\omega}_{\hat{\alpha},i}'(Y - \bar{X}) \xrightarrow{P} A_{i,y}.$$

where $A_{i,y} = \lim_{p\to\infty} p^{-1}\boldsymbol{\mu}_i' P_{\mathcal{U}_m^\perp}(\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}})$ and $\bar{\boldsymbol{\mu}} = n^{-1}\sum_{k=1}^{K} n_k\boldsymbol{\mu}_k$.

(ii) For $i = 1, \ldots, K-1$ and $j = 1, \ldots, K$,

$$\frac{1}{\sqrt{p}}\boldsymbol{\omega}_{\hat{\alpha},i}'(\bar{X}_j - \bar{X}) \xrightarrow{P} \begin{cases} A_{i,j} + n_j^{-1}\tau^2, & \text{if } j = i, \\ A_{i,j} - n_K^{-1}\tau^2, & \text{if } j = K, \\ A_{i,j}, & o.w. \end{cases}$$

Then for $i = 1, \ldots, K-1$ and $j = 1, \ldots, K$,

$$\frac{1}{\sqrt{p}}\boldsymbol{\omega}_{\hat{\alpha},i}'(Y - \bar{X}_j) \xrightarrow{P} \begin{cases} A_{i,yj} - n_j^{-1}\tau^2, & \text{if } j = i, \\ A_{i,yj} + n_K^{-1}\tau^2, & \text{if } j = K, \\ A_{i,yj}, & o.w \end{cases}$$

as $p \to \infty$ where $A_{i,yj} = \lim_{p\to\infty} p^{-1}\boldsymbol{\mu}_i' P_{\mathcal{U}_m^\perp}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_j)$ and Lemma 3 follows. To show (i), note that

$$\frac{1}{\sqrt{p}}\boldsymbol{\omega}'_{\hat{\alpha},i}(Y-\bar{X}) = \sum_{k=1}^{m}\frac{\hat{\alpha}}{\hat{\lambda}_k/p+\hat{\alpha}}\left(\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k{}'\mathbf{d}_i\right)\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k{}'(Y-\bar{X}) + \frac{1}{p}\mathbf{d}_i'(\mathbf{I}_p - \hat{\mathbf{U}}_W\hat{\mathbf{U}}_W')(Y-\bar{X}).$$

From Lemma 4 (ii) and Lemma 5 (i), we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k{}'(Y-\bar{X}) = \frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k{}'\left\{\mathbf{U}\boldsymbol{\Lambda}^{1/2}\left(\boldsymbol{\zeta}-\frac{1}{n}\mathbf{Z}\mathbf{1}_n\right)+\boldsymbol{\mu}_y-\bar{\boldsymbol{\mu}}\right\}\\
&= \sum_{l=1}^{m}(\hat{\mathbf{u}}_k{}'\mathbf{u}_l)\sigma_l(\zeta_l-\bar{\mathbf{z}}_l) + \frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k{}'\boldsymbol{\mu}_y - \sum_{\iota=1}^{K}\frac{n_\iota}{n}\left(\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k'\boldsymbol{\mu}_\iota\right)+o_p(1)\\
&\xrightarrow{P}\sqrt{\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi})+\tau^2}}\sum_{l=1}^{m}\left\{\sigma_l(\zeta_l-\bar{\mathbf{z}}_l)+\cos\theta_{l,y}\delta_y-\sum_{\iota=1}^{K}\frac{n_\iota}{n}\cos\theta_{l,\iota}\delta_\iota\right\}v_{kl}(\boldsymbol{\Phi})
\end{aligned} \tag{A16}$$

as $p\to\infty$ where we use the convention that $\cos\theta_{l,K}=0$ and $\delta_K=0$. Similarly,

$$\begin{aligned}
&\frac{1}{p}\mathbf{d}_i'(Y-\bar{X}) = \frac{1}{p}\left(\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{Z}\mathbf{a}_i+\boldsymbol{\mu}_i\right)'\left\{\mathbf{U}\boldsymbol{\Lambda}^{1/2}\left(\boldsymbol{\zeta}-\frac{1}{n}\mathbf{Z}\mathbf{1}_n\right)+\boldsymbol{\mu}_y-\bar{\boldsymbol{\mu}}\right\}\xrightarrow{P}\\
&A_{i,y}+\sum_{l=1}^{m}\left\{\sigma_l(\bar{\mathbf{z}}_{i,l}-\bar{\mathbf{z}}_{K,l})+\cos\theta_{l,i}\delta_i\right\}\left\{\sigma_l(\zeta_l-\bar{\mathbf{z}}_l)+\cos\theta_{l,y}\delta_y-\sum_{\iota=1}^{K}\frac{n_\iota}{n}\cos\theta_{l,\iota}\delta_\iota\right\}
\end{aligned} \tag{A17}$$

as $p\to\infty$. Combining Lemma 4, Lemma 5 (ii), (A16) and (A17) gives

$$\begin{aligned}
&\frac{1}{\sqrt{p}}\boldsymbol{\omega}'_{\hat{\alpha},i}(Y-\bar{X})\\
&\xrightarrow{P}\sum_{k=1}^{m}\frac{-\tau^2}{\varphi_k(\boldsymbol{\Phi})}\left[\sum_{l'=1}^{m}\left\{\sigma_{l'}(\bar{\mathbf{z}}_{i,l'}-\bar{\mathbf{z}}_{K,l'})+\cos\theta_{l',i}\delta_i\right\}\sqrt{\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi})+\tau^2}}v_{kl'}(\boldsymbol{\Phi})\right]\\
&\times\sqrt{\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi})+\tau^2}}\sum_{l=1}^{m}\left\{\sigma_l(\zeta_l-\bar{\mathbf{z}}_l)+\cos\theta_{l,y}\delta_y-\sum_{\iota=1}^{K}\frac{n_\iota}{n}\cos\theta_{l,\iota}\delta_\iota\right\}v_{kl}(\boldsymbol{\Phi})\\
&+\sum_{l=1}^{m}\left\{\sigma_l(\bar{\mathbf{z}}_{i,l}-\bar{\mathbf{z}}_{K,l})+\cos\theta_{l,i}\delta_i\right\}\left\{\sigma_l(\zeta_l-\bar{\mathbf{z}}_l)+\cos\theta_{l,y}\delta_y-\sum_{\iota=1}^{K}\frac{n_\iota}{n}\cos\theta_{l,\iota}\delta_\iota\right\}\\
&+A_{i,y}-\sum_{k=1}^{m}\sum_{l'=1}^{m}\left\{\sigma_{l'}(\bar{\mathbf{z}}_{i,l'}-\bar{\mathbf{z}}_{K,l'})+\cos\theta_{l',i}\delta_i\right\}\sqrt{\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi})+\tau^2}}v_{kl'}(\boldsymbol{\Phi})\\
&\times\sqrt{\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi})+\tau^2}}\sum_{l=1}^{m}\left\{\sigma_l(\zeta_l-\bar{\mathbf{z}}_l)+\cos\theta_{l,y}\delta_y-\sum_{\iota=1}^{K}\frac{n_\iota}{n}\cos\theta_{l,\iota}\delta_\iota\right\}v_{kl}(\boldsymbol{\Phi})\\
&=A_{i,y}
\end{aligned}$$

as $p\to\infty$ due to the fact that $\sum_{k=1}^{m}v_{kl}(\boldsymbol{\Phi})v_{kl'}(\boldsymbol{\Phi})=I_{l,l'}$ where $I_{l,l'}=1$ if $l=l'$ and $I_{l,l'}=0$ if $l\neq l'$. Next, to show (ii), let $\mathbf{c}_j=(c_{1j},\ldots,c_{nj})'$ where

$$c_{ij} = \begin{cases} 1/n_j - 1/n & \text{if } i \in (\sum_{k=1}^{j-1} n_k, \sum_{k=1}^{j} n_k] \\ -1/n & o.w. \end{cases}$$

Using similar arguments to (i), we obtain

$$\begin{aligned} &\frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k'(\bar{X}_j - \bar{X}) = \frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k'\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{Z}\mathbf{c}_j + \frac{1}{\sqrt{p}}\hat{\mathbf{u}}_k'(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}) \\ &\xrightarrow{P} \sqrt{\frac{\varphi_k(\boldsymbol{\Phi})}{\varphi_k(\boldsymbol{\Phi}) + \tau^2}} \sum_{l=1}^{m} \left\{ \sigma_l(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_l) + \cos\theta_{l,j}\delta_j - \sum_{\iota=1}^{K} \frac{n_\iota}{n} \cos\theta_{l,\iota}\delta_\iota \right\} v_{kl}(\boldsymbol{\Phi}) \end{aligned} \tag{A18}$$

and

$$\begin{aligned} &\frac{1}{p}\mathbf{d}_i'(\bar{X}_j - \bar{X}) = \frac{1}{p}\left(\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{Z}\mathbf{a}_i + \boldsymbol{\mu}_i\right)'\left(\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{Z}\mathbf{c}_j + \boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}\right) \\ &\xrightarrow{P} \sum_{l=1}^{m} \left\{\sigma_l(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l}) + \cos\theta_{l,i}\delta_i\right\} \left\{ \sigma_l(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_l) + \cos\theta_{l,j}\delta_j - \sum_{\iota=1}^{K} \frac{n_\iota}{n} \cos\theta_{l,\iota}\delta_\iota \right\} \\ &+ A_{i,j} + \frac{\tau^2}{n_i} I_{j,i} - \frac{\tau^2}{n_K} I_{j,K} \end{aligned} \tag{A19}$$

as $p \to \infty$. Here, we used the fact that

$$\mathbf{a}_i'\left(\frac{1}{p}\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z}\right)\mathbf{c}_j \xrightarrow{P} \mathbf{a}_i'\left(\mathbf{W}\mathbf{W}' + \tau^2\mathbf{I}_n\right)\mathbf{c}_j = \sum_{l=1}^{m} \sigma_l^2(\bar{\mathbf{z}}_{i,l} - \bar{\mathbf{z}}_{K,l})(\bar{\mathbf{z}}_{j,l} - \bar{\mathbf{z}}_l) + \frac{\tau^2}{n_i} I_{j,i} - \frac{\tau^2}{n_K} I_{j,K}$$

as $p \to \infty$. Then combining (A18), (A19), Lemma 4 and 5 gives (ii). □

Proof of Theorem 4

Proof The classification error rate of $\phi_{PRS-BCNC,\alpha}$ is

$$\mathbb{P}\{\phi_{PRS-BCNC,\alpha}(Y;\mathcal{X}) \neq \pi(Y)\} = 1 - \sum_{y=1}^{K} \mathbb{P}\{\phi_{PRS-BCNC,\alpha}(Y;\mathcal{X}) = y | \pi(Y) = y\}\mathbb{P}\{\pi(Y) = y\}.$$

For any independent observation $Y \in \mathcal{Y}$ with $\pi(Y) = y$, Lemma 3 implies that

$$\mathbf{g}_{\hat{\alpha},j}(Y;\mathcal{X}) \xrightarrow{P} \begin{cases} \mathbf{0}_{K-1}, & j = y, \\ \boldsymbol{\delta}(y,j), & j \neq y \end{cases}$$

as $p \to \infty$ where $\boldsymbol{\delta}(y,j) = (A_{1,yj}, \ldots, A_{K-1,yj})' \in \mathbb{R}^{K-1}$. Note that $\boldsymbol{\delta}(y,j) \neq \mathbf{0}_{K-1}$ by Assumption 3. Hence,

$$\left\|\mathbf{g}_{\hat{\alpha},j}(Y;\mathcal{X})\right\|_2 \xrightarrow{P} \begin{cases} 0, & j = y, \\ \|\boldsymbol{\delta}(y,j)\|_2 > 0, & j \neq y \end{cases}$$

as $p \to \infty$ and $\mathbb{P}\{\phi_{PRS-BCNC,\hat{\alpha}}(Y;\mathcal{X}) = y | \pi(Y) = y\} \to 1$ as $p \to \infty$ for all $y = 1, \ldots, K$. Therefore, $\mathbb{P}\{\phi_{PRS-BCNC,\hat{\alpha}}(Y;\mathcal{X}) \neq \pi(Y)\} \to 0$ as $p \to \infty$. □

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2002256, 2021R1A2C1093526, 2022M3J6A1063021, RS-2023-00218231)

Funding Open Access funding enabled and organized by Seoul National University.

Data availability The MNIST dataset is publicly available; see LeCun et al. (2010).

References

Ahn, J., & Jeon, Y. (2015). Sparse HDLSS discrimination with constrained data piling. *Computational Statistics & Data Analysis, 90*, 74–83. https://doi.org/10.1016/j.csda.2015.04.006

Ahn, J., Lee, M. H., & Lee, J. A. (2019). Distance-based outlier detection for high dimension, low sample size data. *Journal of Applied Statistics, 46*(1), 13–29. https://doi.org/10.1080/02664763.2018.1452901

Ahn, J., Lee, M. H., & Yoon, Y. J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica, 22*(2), 443–464. https://doi.org/10.5705/ss.2010.148

Ahn, J., & Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika, 97*(1), 254–259. https://doi.org/10.1093/biomet/asp084

Ahn, J., Marron, J. S., Muller, K. M., et al. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika, 94*(3), 760–766. https://doi.org/10.1093/biomet/asm050

Ardeshir N, Sanford C, & Hsu DJ (2021) Support vector machines and linear regression coincide with very high-dimensional features. In: Advances in Neural Information Processing Systems, 4907–4918

Bartlett, P. L., Long, P. M., Lugosi, G., et al. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences, 117*(48), 30063–30070. https://doi.org/10.1073/pnas.1907378117

Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numerica, 30*, 87–201. https://doi.org/10.1017/s0962492921000027

Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica, 30*, 203–248. https://doi.org/10.1017/s0962492921000039

Belkin, M., Hsu, D., Ma, S., et al. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Belkin, M., Hsu, D., & Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science, 2*(4), 1167–1180. https://doi.org/10.1137/20M1336072

Bogachev VI (2007) Measure Theory, vol 1. Springer Science & Business Media

Cao Y, Gu Q, & Belkin M (2021) Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In: Advances in Neural Information Processing Systems, 8407–8418

Chang, W., Ahn, J., & Jung, S. (2021). Double data piling leads to perfect classification. *Electronic Journal of Statistics, 15*(2), 6382–6428. https://doi.org/10.1214/21-ejs1945

Chatterji, N. S., & Long, P. M. (2021). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research, 22*(129), 1–30.

Chung, H. C., & Ahn, J. (2021). Subspace rotations for high-dimensional outlier detection. *Journal of Multivariate Analysis, 183*, 104713. https://doi.org/10.1016/j.jmva.2020.104713

Dai, X., Müller, H. G., & Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika, 104*(3), 545–560. https://doi.org/10.1093/biomet/asx024

Delaigle, A., & Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74*(2), 267–286. https://doi.org/10.1111/j.1467-9868.2011.01003.x

Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(3), 427–444. https://doi.org/10.1111/j.1467-9868.2005.00510.x

Hastie, T., Montanari, A., Rosset, S., et al. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics, 50*(2), 949–986. https://doi.org/10.1214/21-AOS2133

Horn, R. A., & Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press.

Hsu, D., Muthukumar, V., & Xu, J. (2022). On the proliferation of support vectors in high dimensions. *Journal of Statistical Mechanics: Theory and Experiment, 11*, 114011. https://doi.org/10.1088/1742-5468/ac98a9

Huang, H., Liu, Y., Du, Y., et al. (2013). Multiclass distance-weighted discrimination. *Journal of Computational and Graphical Statistics, 22*(4), 953–969. https://doi.org/10.1080/10618600.2012.700878

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics, 29*(2), 295–327. https://doi.org/10.1214/aos/1009210544

Jung, S. (2018). Continuum directions for supervised dimension reduction. *Computational Statistics & Data Analysis, 125*, 27–43. https://doi.org/10.1016/j.csda.2018.03.015

Jung, S. (2022). Adjusting systematic bias in high dimensional principal component scores. *Statistica Sinica, 32*, 939–959. https://doi.org/10.5705/ss.202019.0400

Jung, S., & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics, 37*(6B), 4104–4130. https://doi.org/10.1214/09-AOS709

Jung, S., Sen, A., & Marron, J. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis, 109*, 190–203. https://doi.org/10.1016/j.jmva.2012.03.005

Kim T, Ahn J, & Jung S (2022) Double data piling for heterogeneous covariance models. arXiv preprint arXiv:2211.15562

Kobak, D., Lomond, J., & Sanchez, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research, 21*(169), 1–16.

Kolmogorov, A. N., & Rozanov, Y. A. (1960). On strong mixing conditions for stationary gaussian processes. *Theory of Probability & its Applications, 5*(2), 204–208. https://doi.org/10.1137/1105018

LeCun Y, Cortes C, & Burges C (2010) MNIST handwritten digit database. ATT Labs [Online] 2. http://yann.lecun.com/exdb/mnist

Lee, M. H., Ahn, J., & Jeon, Y. (2013). Hdlss discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics, 22*(2), 433–451. https://doi.org/10.1080/10618600.2012.681235

Mahdaviyeh Y, & Naulet Z (2020) Risk of the least squares minimum norm estimator under the spike covariance model. arXiv preprint arXiv:1912.13421https://doi.org/10.48550/arXiv.1912.13421

Marron, J. S., Todd, M. J., & Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association, 102*(480), 1267–1271. https://doi.org/10.1198/016214507000001120

Mei, S., & Montanari, A. (2021). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics, 75*(4), 667–766. https://doi.org/10.1002/cpa.22008

Montanari A, Ruan F, & Sohn Y, et al (2023) The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. arXiv preprint arXiv:1911.01544https://doi.org/10.48550/arXiv.1911.01544

Muthukumar, V., Narang, A., Subramanian, V., et al. (2021). Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research, 22*(222), 1–69.

Muthukumar, V., Vodrahalli, K., Subramanian, V., et al. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory, 1*(1), 67–83. https://doi.org/10.1109/jsait.2020.2984716

Qiao X, & Zhang L (2015) Distance-weighted support vector machine. arXiv preprint arXiv:1310.3003https://doi.org/10.48550/arXiv.1310.3003

Qiao, X., Zhang, H. H., Liu, Y., et al. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association, 105*(489), 401–414. https://doi.org/10.1198/jasa.2010.tm08487

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

Stewart, G. W., & Sun, J. G. (1990). *Matrix Perturbation Theory*. Academic Press.

Tsigler A, & Bartlett PL (2020) Benign overfitting in ridge regression. arXiv preprint arXiv:2009.14286https://doi.org/10.48550/ARXIV.2009.14286

Wang, W., & Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics, 45*(3), 1342–1374. https://doi.org/10.1214/16-AOS1487

Wang K, Muthukumar V, & Thrampoulidis C (2021) Benign overfitting in multiclass classification: All roads lead to interpolation. In: Advances in Neural Information Processing Systems, pp 24164–24179

Wang, K., & Thrampoulidis, C. (2022). Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science, 4*(1), 260–284. https://doi.org/10.1137/21m1415121

Wu D, & Xu J (2020) On the optimal weighted l_2 regularization in overparameterized linear regression. In: Advances in Neural Information Processing Systems, pp 10112–10123

Xue, K., Yang, J., & Yao, F. (2023). Optimal linear discriminant analysis for high-dimensional functional data. *Journal of the American Statistical Association, 118*, 1–10. https://doi.org/10.1080/01621459.2022.2164288

Zhang C, Bengio S, & Hardt M, et al (2017) Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530https://doi.org/10.48550/arxiv.1611.03530

Authors and Affiliations

Taehyun Kim[1,2] · Woonyoung Chang[1,3] · Jeongyoun Ahn[4] · Sungkyu Jung[1]

✉ Sungkyu Jung
sungkyu@snu.ac.kr

Taehyun Kim
tk3036@columbia.edu

Woonyoung Chang
woonyouc@andrew.cmu.edu

Jeongyoun Ahn
jyahn@kaist.ac.kr

[1] Department of Statistics, Seoul National University, Seoul, South Korea

[2] Department of Statistics, Columbia University, New York, New York, USA

[3] Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[4] Department of Industrial and Systems Engineering, KAIST, Daejeon, South Korea