Electronic Journal of Statistics
Vol. 15 (2021) 6382–6428
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1945

Double data piling leads to perfect classification[*]

Woonyoung Chang

Department of Statistics and Data Science, Carnegie Mellon University,
Pittsburgh, PA 15213, United States
e-mail: woonyouc@andrew.cmu.edu

Jeongyoun Ahn

Department of Industrial and Systems Engineering, KAIST,
Daejeon 34141, South Korea
e-mail: jyahn@kaist.ac.kr

and

Sungkyu Jung

Department of Statistics, Seoul National University,
Seoul 08826, South Korea
e-mail: sungkyu@snu.ac.kr

Abstract: Data piling refers to the phenomenon that training data vectors from each class project to a single point for classification. While this interesting phenomenon has been a key to understanding many distinctive properties of high-dimensional discrimination, the theoretical underpinning of data piling is far from properly established. In this work, high-dimensional asymptotics of data piling is investigated under a spiked covariance model, which reveals its close connection to the well-known ridged linear classifier. In particular, by projecting the ridge discriminant vector onto the subspace spanned by the leading sample principal component directions and the maximal data piling vector, we show that a negatively ridged discriminant vector can asymptotically achieve data piling of independent test data, essentially yielding a perfect classification. The second data piling direction is obtained purely from training data and shown to have a maximal property. Furthermore, asymptotic perfect classification occurs only along the second data piling direction.

MSC2020 subject classifications: Primary 62H30, 62H25; secondary 62J07.
Keywords and phrases: Discrimination, high dimension low sample size, maximal data piling, negative ridge, spike covariance model.

Received July 2021.

Contents

[*] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2002256, 2021R1A2C1093526).

1. Introduction

Classification of high dimensional data has become an extremely common statistical problem. For a two-group classification, we use the high-dimension, low-sample-size (HDLSS) asymptotic regime (Hall, Marron and Neeman, 2005), in which the dimension p increases while the sample size n is fixed, to reveal a perhaps counter-intuitive phenomenon in classification of HDLSS data.

Consider a situation where a classification rule is given by a linear separating hyperplane, or equivalently via the orthogonal projection of data onto its normal vector. A classical choice for such a vector is the Fisher's linear discriminant vector w_{FLD} (Fisher, 1936), which maximizes the between-group scatter $(w^T d)^2$ while minimizing the within-group scatter $w^T S w$, where $d = \bar{X}_1 - \bar{X}_2$ and S is the $p \times p$ pooled covariance matrix of rank $\min\{p, n-2\}$. While $w_{\text{FLD}} = S^{-1} d$

is not defined for high-dimensional data with $p > n - 2$, a natural extension is the *maximal data piling* direction vector (Ahn and Marron, 2010)

$$w_{\mathrm{MDP}} = \underset{w:\|w\|=1}{\operatorname{argmax}} (w^T d)^2 \quad \text{subject to} \quad w^T S w = 0.$$

The term data-piling in the binary classification setting refers to the phenomenon that all data are projected to exactly two points, one for each group. Among such vectors exhibiting data piling, w_{MDP} uniquely maximizes the between-group scatter. This "first" data-piling is observed for any given data whenever $p > n - 2$, and may be viewed as a sign of overfitting; independent test data are not projected to the same piling locations.

In this paper, we reveal that there exists a "second" data piling vector onto which any independent observations are projected to two distinct points, asymptotically. This second data piling vector can be obtained purely from the training data. Thus, when used for classification, the second data piling direction leads to an asymptotic perfect classification.

We develop the second data-piling in a dense signal setting for which the true mean difference exists in almost all coordinates, i.e., $\|\mu_1 - \mu_2\| = O(p^{1/2})$, and using a spiked model for the common covariance Σ. In particular, the spiked covariance model with m spikes, for a fixed $m \geq 0$, assumes that, for a given $\beta \in [0, 1]$, the m leading eigenvalues of Σ increases at the order of p^β, while the rest of eigenvalues are nearly constant. This model has been commonly used in the high-dimensional asymptotic studies for principal component analysis and factor models; see Jung, Lee and Ahn (2018); Hellton and Thoresen (2017); Fan et al. (2021) and references therein. Most of the previous HDLSS-asymptotic studies on classification (Hall, Marron and Neeman, 2005; Qiao et al., 2010; Jung, 2018) are limited to the simple null case of $\beta = 0$, i.e., Σ is the scaled identity matrix. Allowing $\beta \in [0, 1)$ also results in a similar conclusion (Yata and Aoshima, 2020). Our findings are obtained under the more interesting and realistic situation at $\beta = 1$, requiring the variables to be meaningfully correlated with each other. To the best of our knowledge, literature on the theoretical research on classification under this scenario is scarce. An exception is the work of Aoshima and Yata (2019), in which the authors proposed to "remove" the leading eigenspace for better classification performances. Relation of our findings to Aoshima and Yata (2019) is further discussed in Section 5.

The first and second data piling direction vectors turn out to be closely related to the ridged linear discriminant vector $w_\alpha = (S + \alpha I_p)^{-1} d$, where I_p is the $p \times p$ identity matrix (Di Pillo, 1976), for which both positive and negative values for the ridge parameter α are considered. The first maximal-data-piling direction w_{MDP} is the direction of the "ridgeless" vector $\lim_{\alpha \downarrow 0} w_\alpha$. Since our model suggests that only the first m leading eigenvectors $\hat{u}_1, \ldots, \hat{u}_m$ of S are meaningful estimates of their population counterparts, we define a projected ridge discriminant direction v_α, given by projecting w_α onto $\mathrm{span}(\hat{u}_1, \ldots, \hat{u}_m, w_{\mathrm{MDP}})$ for most values of $\alpha \in \mathbb{R}$; see Section 2 for a precise definition of v_α. We show that the second data piling occurs at $v_{\hat{\alpha}}$, for an $\hat{\alpha}$ strictly negative, but not at any other choices of α. This implies that an asymptotic perfect classification can happen only at $v_{\hat{\alpha}}$.

Note that the second data piling direction is found by allowing the ridge parameter to be negative. Recently, for over-parameterized regression models with $p > n$, Kobak, Lomond and Sanchez (2020); Wu and Xu (2020); Tsigler and Bartlett (2020) have pointed out that negatively ridged coefficient estimators can be optimal. In particular, Tsigler and Bartlett (2020) have shown that a spiked covariance model with the number of spikes much smaller than the sample size is a key for a negative ridge parameter to be optimal in ridge regression. Our finding with the m-spiked model parallels this observation, in a classification setting. In the linear regression context, extreme overfitting of ridgeless estimators is found to perform surprisingly well (Hastie et al., 2019; Holzmüller, 2020; Bartlett et al., 2020), a phenomenon called "double descent". The ridgeless estimator in the classification context is exactly the first data piling vector w_{MDP}. Our results further suggest that it is possible to asymptotically interpolate test data by the second data piling vector. We also note that an asymptotic perfect classification of functional data is shown to be possible (Delaigle and Hall, 2012). There, a key condition enabling perfect classification is that the norm of the mean difference is comparable to or larger than the standard deviation of the leading principal component scores, which is analogous to our model with $\|\mu_1 - \mu_2\| = O(p^{1/2})$, $\lambda_i^{1/2} = O(p^{\beta/2})$ for $i = 1, \ldots, m$ and $\beta \leq 1$.

The HDLSS asymptotic regime we adopt has been used to reveal some unique characteristics of HDLSS data. Recent developments on the HDLSS asymptotic studies of various multivariate methods are surveyed by Aoshima et al. (2018).

2. Linear discriminant directions in high dimensions

2.1. Negatively ridged discriminant directions

For $i = 1, 2$, let $X_{i1}, \ldots, X_{in_i} \in \mathbb{R}^p$ denote a sample drawn from an absolutely continuous distribution on $\mathbb{R}^p$ with mean μ_i and common covariance matrix Σ. We assume $p > n := n_1 + n_2$. Let $d = \bar{X}_1 - \bar{X}_2$ be the vector of sample mean difference, and $S = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T / n$ be the sample within-group covariance matrix. We consider a ridge-type discriminant vector: For a ridge-parameter $\alpha \in (-\infty, \infty)$,

$$\tilde{w}_\alpha = \alpha_p \left(S + \alpha_p I_p\right)^{-1} d, \tag{2.1}$$

where $\alpha_p = \alpha p$. We write $w_\alpha = \tilde{w}_\alpha / \|\tilde{w}_\alpha\|$. Allowing α_p to increase along p will be convenient in our analysis of diverging p. In (2.1), we have extended the range of ridge parameter to include negative real values. This is in contrast to the conventional range $\alpha \in [0, \infty)$ considered in Friedman (1989); Guo, Hastie and Tibshirani (2007); Lee, Ahn and Jeon (2013) and many others. The price we pay is that (2.1) is in fact ill-defined for some $\alpha \in (-\infty, 0]$ with rank-deficient $S + \alpha_p I_p$. In what follows, we precisely redefine w_α for $\alpha \in [-\infty, \infty]$, by filling in the ill-defined locations.

Let $S = \hat{U}_1 \hat{\Lambda} \hat{U}_1^T$ be the eigen-decomposition of S such that $\hat{U}_1 = [\hat{u}_1, \ldots, \hat{u}_{n-2}]$ consists of the orthogonal eigenvectors, and $\hat{\Lambda}$ be the diagonal matrix whose ith entry $\hat{\lambda}_i$ $(i = 1, \ldots, n-2)$ is the ith largest eigenvalue. (There are exactly $n-2$ positive eigenvalues almost surely.) Let $\hat{U}_2$ be a $p \times (p - n + 2)$ matrix whose columns form an orthonormal basis for the nullspace of S. Then, for α at which (2.1) is defined, we decompose $\tilde{w}_\alpha$ into three parts:

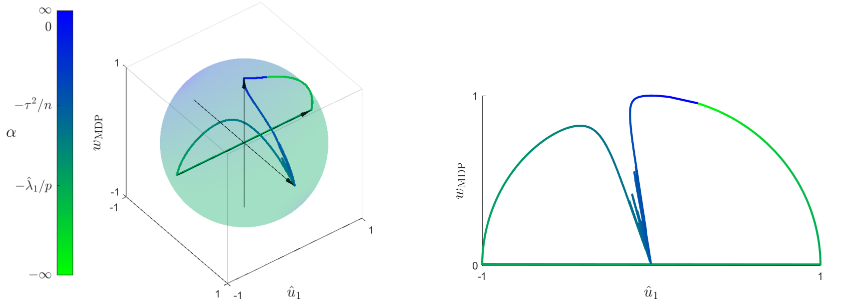
$$\begin{aligned}\tilde{w}_\alpha &= \alpha_p (S + \alpha_p I_p)^{-1} d = \alpha_p \hat{U}_1 \left(\hat{\Lambda} + \alpha_p I\right)^{-1} \hat{U}_1^T d + \hat{U}_2 \hat{U}_2^T d \\ &= \sum_{i=1}^{m} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d + \sum_{i=m+1}^{n-2} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d + \hat{U}_2 \hat{U}_2^T d. \end{aligned} \tag{2.2}$$

The decomposition above reveals that w_α as well as $\tilde{w}_\alpha$ are decomposed into the weighted sum of projections of the mean difference d onto three orthogonal subspaces. The first two subspaces belong to the column space of S. We choose for the first m sample eigenvectors $\hat{u}_1, \ldots, \hat{u}_m$ to constitute the first subspace, in the anticipation that the role of the leading m eigenvectors is distinct from the rest of eigenvectors under the m-component model we consider. The last subspace, spanned by $\hat{U}_2$, is the nullspace of S.

The set of α at which (2.2) is not defined is $\{0, \pm\infty\} \cup \{-\hat{\lambda}_i/p : i = 1, \ldots, n-2\}$. It can be seen that the last term of (2.2) is parallel to the maximal data piling direction, and that $\lim_{\alpha \to 0} w_\alpha = w_{\mathrm{MDP}}$. Similarly, we have $\lim_{\alpha \to \pm\infty} w_\alpha = d/\|d\|$ and $\lim_{\alpha \uparrow -\hat{\lambda}_i/p} w_\alpha = \hat{u}_i$, $\lim_{\alpha \downarrow -\hat{\lambda}_i/p} w_\alpha = -\hat{u}_i$, $(i = 1, \ldots, n-2)$. We set $w_0 := w_{\mathrm{MDP}}$, $w_{\pm\infty} := d/\|d\|$, and $w_{-\hat{\lambda}_i/p} := \hat{u}_i$, so that w_α is left-continuous at $-\hat{\lambda}_i/p$ and continuous elsewhere. Then, the discontinuities of the parameterization path $\alpha \mapsto w_\alpha$ are exactly the $n-2$ sign flips, each of which occurs when $-\alpha_p$ crosses an eigenvalue of S.

It helps to visualize the modified w_α as a trajectory along varying α. For this, temporarily assume a one-component model, i.e., $m = 1$. The parameterization path $\alpha \mapsto w_\alpha$ is visualized in Fig. 1 for a high-dimensional data with $p = 5000$ and $n_1 = n_2 = 20$ (the model for this data set is decribed in Section 4.1). To visualize the p-vector w_α in a three-dimensional plot, we make use of the decomposition (2.2). For any $\alpha \in [-\infty, \infty]$, the ridged discriminant vector w_α lies in the $(n-1)$-dimensional subspace $\mathcal{S}_X$ spanned by $\{X_{ij} - \bar{X}\}$, where $\bar{X} = \sum_i \sum_j X_{ij}/n$, or equivalently, by the column space of S and d. Recall that $m = 1$ in this example. The first and third terms of (2.2) are spanned by $\hat{u}_1, w_{\mathrm{MDP}} \in \mathcal{S}_X$, respectively, which provide the two axes of the figure. The remaining $n-3$ dimensions of $\mathcal{S}_X$, corresponding to the second term of (2.2), are collapsed to a nonnegative value. For the data set used in Fig. 1, it is apparent that, for most values of α, w_α is close to the plane spanned by $\hat{u}_1$ and w_{MDP}. It turns out that the subspace $\mathcal{S} = \mathrm{span}(\hat{u}_1, w_{\mathrm{MDP}})$ is the only meaningful subspace for large p, while the nullspace of $\mathcal{S}$, within $\mathcal{S}_X$, does not possess any discriminative information. We will show this more carefully in the next subsection. For general $m \geq 1$, the subspace of interest is

$$\mathcal{S} = \mathrm{span}(\hat{u}_1, \ldots, \hat{u}_m, w_{\mathrm{MDP}}). \tag{2.3}$$



(a) (b)

FIG 1. *The parameterization path of w_α lies close to $\mathcal{S}$ for most α. (a) Shown are the path $\alpha \mapsto (w_\alpha^T \hat{u}_1, b_\alpha, w_\alpha^T w_{\mathrm{MDP}})$, where $b_\alpha = \|\tilde{w}^{(2)}\|/\|\tilde{w}_\alpha\| \geq 0$, and $\tilde{w}^{(2)}$ is the second term of (2.2). w_0 is the maximal data piling direction, and $w_{\pm\infty}$ is the mean difference direction. As α crosses $-\hat{\lambda}_1/p$, w_α is flipped. (b) The path of w_α projected onto $\mathcal{S} = \mathrm{span}(\hat{u}_1, w_{\mathrm{MDP}})$. See Appendix B for details on the second axis and b_α to see how the $n-3$ sign flips are suppressed for visualization.*

2.2. Concentration of ridge directions in high dimensions

Key conditions are described with respect to the mean difference $\mu = \mu_1 - \mu_2$ and the principal component structure of the common covariance matrix Σ. Denote the eigen-decomposition of Σ by $\Sigma = U\Lambda U^T = \sum_{i=1}^p \lambda_i u_i u_i^T$, where $U = [u_1, \ldots, u_p]$ collects the eigenvectors, and the entries of the diagonal matrix Λ are the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$. Assumptions 1–3 play important roles in describing the high-dimensional asymptotic behaviors of w_α and associated classifiers.

Assumption 1. *For $m \geq 1$, $\sigma_i^2, \tau_i^2 > 0$ and $0 \leq \beta \leq 1$, the eigenvalues of the covariance matrix Σ are $\lambda_i = p^\beta \sigma_i^2 + \tau_i^2$ $(i = 1, \ldots, m)$ and $\lambda_i = \tau_i^2$ $(i = m+1, \ldots, p)$, where $\max \tau_i^2$ is uniformly bounded, and $\sum_{i=1}^p \tau_i^2/p \to \tau^2$ as $p \to \infty$ for some $\tau^2 > 0$.*

Remark 2.1. *A seemingly more relaxed model may be assumed in place of Assumption 1. For example, for some natural number M, we may set $1 \geq \beta_1 \geq \beta_2 \geq \cdots \geq \beta_M \geq 0$ and $\lambda_i = O(p^{\beta_i})$ for $i = 1, \ldots, M$. This model is asymptotically equivalent to the model assumed in Assumption 1 with $m = \#\{\beta_i = 1\}$, due to the following reason: There is no difference between $\lambda_i = O(p^{\beta_i})$ with $\beta_i < 1$ and $\lambda_i = O(1)$ in the HDLSS asymptotic results we derive; see, e.g., Lemma 3.6. Note that allowing $\beta_i \in (0, 1)$ for $i > m$ should make a difference in the convergence rates of our conclusions. Since we do not explicitly state the rates of convergence, we use Assumption 1 for the sake of simplicity.*

Assumption 2. *There exists $\delta \in (0, \infty)$ such that $\|\mu\|^2/p \to \delta^2$ as $p \to \infty$.*

For a vector $w \in \mathbb{R}^p$ and a subspace $\mathcal{A}$ of $\mathbb{R}^p$, we write $P_{\mathcal{A}}$ and $P_{\mathcal{A}} w$ for the orthogonal projection matrix and the orthogonal projection of w, respectively, onto $\mathcal{A}$. Let $\mathcal{U}_m = \operatorname{span}(u_1, \ldots, u_m)$ be the subspace formed by the first m leading eigenvectors.

Assumption 3. *There exists $k \in [0, 1)$ such that $\|P_{\mathcal{U}_m}\mu\|/\|\mu\| \to k$ as $p \to \infty$. Moreover, there exists $k_i \in [0, 1)$ such that $u_i^T\mu/\|\mu\| \to k_i$ as $p \to \infty$ for $i = 1, \ldots, m$.*

The m-component model in Assumption 1 is routinely assumed in classification problems (Qiao et al., 2010; Ahn and Marron, 2010; Yata and Aoshima, 2012) and is a special case of spiked covariance models (Hellton and Thoresen, 2017; Jung, Lee and Ahn, 2018; Ishii, Yata and Aoshima, 2019; Fan et al., 2021). The parameter $\beta \in [0, 1]$ determines the order of magnitude of the m leading eigenvalues $\lambda_1, \ldots, \lambda_m$ of Σ. The asymptotic result obtained by assuming $\beta \in (0, 1)$ is equivalent to the simple null case $\beta = 0$ (*cf.* Ahn et al., 2007; Yata and Aoshima, 2020), so we pay a special attention to the $\beta = 1$ case. Note that the case of $\beta = 0$ (or, equivalently, $m = 0$) is easier to classify, since the Mahalanobis distance $\|\Sigma^{-1/2}\mu\|_2$ is strictly larger than that under $\beta = 1$ and $m \geq 1$. Assumption 3 introduces k that controls the asymptotic portion of the mean difference μ along the m-dimensional principal subspace $\mathcal{U}_m$. We do not allow $k = 1$, as in such a case the mean difference vector μ is completely within the subspace $\mathcal{U}_m$ with larger variance. Put differently, the quantity $1 - k^2 > 0$ is the portion of the mean difference in the subspace of relatively smaller magnitude of noise.

The class of distributions we consider is given by the following assumption on the true principal scores $z_{ij} = \Lambda^{-1/2}U^T(X_{ij} - \mu_i)$. These principal scores may not be independent with each other unless we assume normality.

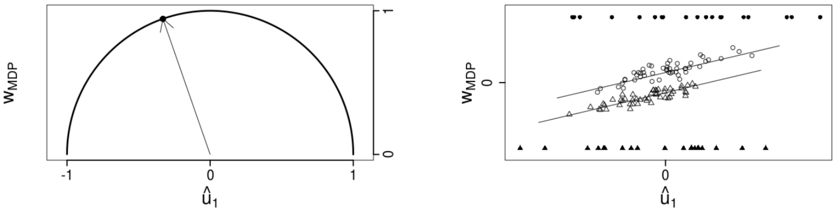
Assumption 4. *The elements of the p-vector z_{ij} have uniformly bounded fourth moments, and for each p, z_{ij} consists of the first p elements of an infinite random sequence $(z_{(1)}, z_{(2)}, \ldots)_{i,j}$, which is ρ-mixing under some permutation.*

One of our main tools of analysis is the law of large numbers applied across variables ($p \to \infty$), rather than across sample ($n \to \infty$). For this, the dependency among the principal scores is controlled by the ρ-mixing condition; see Appendix A for definition. A sequence of independent normally distributed random variables satisfies the ρ-mixing condition. Assumption 4 is much weaker than normality and independence, yet it enables an application of the law of large numbers, as shown in Jung and Marron (2009).

We are now ready to state our result on w_α. We define $\operatorname{Angle}(w, \mathcal{A}) := \cos^{-1}\{|w^T P_{\mathcal{A}} w|/(\|w\|\|P_{\mathcal{A}} w\|)\}$. For two unit vectors $w_1, w_2 \in \mathbb{R}^p$, $\operatorname{Angle}(w_1, w_2) = \cos^{-1}|w_1^T w_2|$ is the acute angle formed by the two vectors. The notation $\xrightarrow{P}$ represents convergence in probability. Throughout, we let n be fixed and $p \to \infty$.

Theorem 2.1. *Suppose Assumptions 1–4 hold and $\beta = 1$ in Assumption 1. For any $\alpha \in \mathbb{R} \setminus \{-\tau^2/n\}$,*

$$\operatorname{Angle}(w_\alpha, \mathcal{S}) \xrightarrow{P} 0, \tag{2.4}$$



(a) (b)

Fig 2. *(a) Diagram illustrating $\mathcal{S}$ and v_α obtained from the data $\mathcal{X}$ used in Fig. 1. The arrow corresponds to v_α at $\alpha = -\tau^2/n$. (b) Projections onto $\mathcal{S}$ of the data $\mathcal{X}$ (group 1: filled circles, group 2: filled triangles) and an independent data set $\mathcal{Y}$ (group 1: circles, group 2: triangles). The origin of the plot is at $\bar{X}$. The independent data $\mathcal{Y}$ are concentrated along the two parallel lines (3.1), which appear to be orthogonal to $v_{-\tau^2/n}$, shown as the arrow in (a).*

as $p \to \infty$.

All proofs are contained in Appendix C.

Theorem 2.1 confirms that the concentration of w_α towards $\mathcal{S}$ for almost all values of α, illustrated in Fig. 1, is bound to happen in high dimensions. The only exception is at $\alpha = -\tau^2/n$, to which all scaled eigenvalues $-\hat{\lambda}_i/p$ ($i = m+1, \ldots, n-2$) converge. (The asymptotic behaviors of eigenvalues of S are stated in Lemma C.2.) Heuristically, at $\alpha = \alpha_i := -\hat{\lambda}_i/p \approx -\tau^2/n$ for $i \geq m+1$, we have $w_{\alpha_i} = \hat{u}_i$, which is orthogonal to $\mathcal{S}$ for every p.

We remark that if the assumption $\beta = 1$ in Theorem 2.1 is replaced by $\beta \in [0,1)$, then a stronger, yet less interesting, statement can be given. Specifically, if $\beta \in [0,1)$, then the role of $\hat{u}_1, \ldots, \hat{u}_m$ is no longer different from the rest of sample principal component directions, and for $\alpha \neq -\tau^2/n$, $\text{Angle}(w_\alpha, w_{\text{MDP}}) \xrightarrow{P} 0$ as $p \to \infty$. This can be shown by an argument used in Jung (2018). In contrast, in the m-component model with $\beta = 1$, $\text{Angle}(w_\alpha, w_{\text{MDP}})$ can be large in the limit, depending on the choice of α, as depicted for finite p in Fig. 1(b).

Based on the above, it will be convenient to consider a projection of w_α onto $\mathcal{S}$, to which almost all ridge directions converge in high dimensions. We propose to use

$$v_\alpha \propto \sum_{i=1}^{m} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d + \hat{U}_2 \hat{U}_2^T d, \tag{2.5}$$

satisfying $\|v_\alpha\| = 1$. This v_α is not exactly the projection $P_{\mathcal{S}} w_\alpha / \|P_{\mathcal{S}} w_\alpha\|$, because $P_{\mathcal{S}} w_\alpha = 0$ at $\alpha = -\hat{\lambda}_i/p$ for any $i = m+1, \ldots, n-2$. The v_α is given by filling in the $n-m-2$ undefined points of the scaled projections. It can be seen that v_α simply ignores the second term of (2.2). For the special case of $m = 1$, the trajectory of the unit vectors v_α over $\alpha \in [-\infty, \infty]$, which is a unit semicircle on the plane $\mathcal{S}$, is illustrated in Fig. 2(a).

3. Double data piling and perfect classification

3.1. First data piling

The data piling in two-group classification problems refers to the phenomenon that when data are projected onto a vector w, the projected data are piled on two points, one for each group (Ahn and Marron, 2010). This "first" data piling can be observed whenever $p > n - 2$. Specifically, for any vector w in the nullspace of S, i.e., $w \in \text{span}(\hat{U}_2)$ in the decomposition (2.2), data piling occurs. Among those, the maximal data piling direction w_{MDP} uniquely maximizes the distance between two piling locations. For example, in Fig. 2(b) the data $\mathcal{X} := \{x_{ij} : i = 1, 2, j = 1, \ldots, n_i\}$ projected to w_{MDP} are piled on two locations on the w_{MDP} axis, and the distance between them is the largest. Note that w_α, v_α and $\mathcal{S}$ depend only on the data set $\mathcal{X}$.

3.2. Second data piling

The second data piling phenomenon occurs for new, independent data $\mathcal{Y}$, sampled from the same model as $\mathcal{X}$ and independent of $\mathcal{X}$. One may regard $\mathcal{X}$ as training data, $\mathcal{Y}$ as testing data. Figure 2(b) plots both $\mathcal{X}$ and $\mathcal{Y}$ of a one-component model, i.e., $m = 1$, projected onto $\mathcal{S}$. While $\mathcal{Y}$ projected on w_{MDP} does not exhibit a data piling, it is interesting to observe that $P_{\mathcal{S}}\mathcal{Y}$ tend to lie on two parallel straight lines, one for each group. It turns out the piling of $\mathcal{Y}$ onto these two lines occurs asymptotically as $p \to \infty$. We call this new phenomenon as the second data piling, this time for independent observations. For the direction vector v on $\mathcal{S}$ orthogonal to both lines, $P_v\mathcal{Y}$ exhibits data piling on two points, asymptotically.

In the following, we show that such direction v can be identified purely from the training data $\mathcal{X}$, even before observing the new data $\mathcal{Y}$. Moreover, we show that such second data piling direction is also maximal in a sense similar to the first maximal data piling.

We begin by focusing on the two lines, l_1 and l_2, in Fig. 2(b). The direction of the parallel lines has a close connection with u_1, the first *true* principal component direction. Let $u_{1,\mathcal{S}} = P_{\mathcal{S}}u_1$ be the projection of u_1 onto $\mathcal{S}$, and write $\kappa_{\text{MDP}} := \|w_{\text{MDP}}^T(\bar{X}_1 - \bar{X}_2)\|/\sqrt{p}$, which is the distance between the two piles on the first data piling direction, scaled by $\sqrt{p}$. The two lines are

$$l_i = \{tu_{1,\mathcal{S}} + \kappa_i w_{\text{MDP}} + P_{\mathcal{S}}\bar{X}/\sqrt{p} : t \in \mathbb{R}\} \quad (i = 1, 2), \tag{3.1}$$

where $\kappa_1 = (1 - n_1/n)(1 - k^2)\delta^2/\kappa_{\text{MDP}}$, and $\kappa_2 = -(1 - n_2/n)(1 - k^2)\delta^2/\kappa_{\text{MDP}}$. The two lines l_1 and l_2 are parallel to $u_{1,\mathcal{S}}$.

For general cases where $m \geq 1$, the two lines (3.1) are naturally extended to two m-dimensional affine subspaces, L_1 and L_2. Let $u_{i,\mathcal{S}} = P_{\mathcal{S}}u_i$ be the projection of u_i onto $\mathcal{S}$ $(i = 1, \ldots, m)$, and write $U_{m,\mathcal{S}} = [u_{1,\mathcal{S}}, \ldots, u_{m,\mathcal{S}}]$. We define

$$L_i = \left\{U_{m,\mathcal{S}}t + \kappa_i w_{\text{MDP}} + P_{\mathcal{S}}\bar{X}/\sqrt{p} : t \in \mathbb{R}^m\right\} \quad (i = 1, 2). \tag{3.2}$$

We now confirm that the independent data are piled along the affine subspaces. For any observation $Y \in \mathcal{Y}$, we write $\pi(Y) = i$ if Y belongs to group i.

Theorem 3.1. *Suppose Assumptions 1–4 hold and $\beta = 1$. For any observation $Y \in \mathcal{Y}$, write $Y_{\mathcal{S}} := P_{\mathcal{S}} Y/\sqrt{p}$. Then, for $i = 1, 2$ and for any $\epsilon > 0$,*

$$\lim_{p\to\infty} \mathbb{P}\{\inf_{a \in L_i} \|Y_{\mathcal{S}} - a\| > \epsilon \mid \pi(Y) = i\} = 0.$$

Theorem 3.1 reveals that exact second data piling for independent data occurs asymptotically. Even for the finite-dimensional example in Fig. 2(b), the distance between $Y_{\mathcal{S}}$ and L_i is small enough to perceive the second data piling.

The parameterization of L_i in (3.2) involves unknown parameters $u_1, \ldots, u_m$, and it is impossible to obtain L_i, or the direction orthogonal to L_i, from only $\mathcal{X}$. Our next result shows that there exists an $\hat{\alpha}$, purely a function of the data $\mathcal{X}$, such that $v_{\hat{\alpha}}$ is a direction asymptotically perpendicular to L_i. As Theorem 3.2 shows, any HDLSS-consistent estimator of $-\tau^2/n$ is such an $\hat{\alpha}$. We say $\hat{\theta}$ is an HDLSS-consistent estimator for θ if for any $\epsilon > 0$, $\lim_{p\to\infty} \mathbb{P}(|\hat{\theta} - \theta| > \epsilon) = 0$.

Theorem 3.2. *Suppose Assumptions 1–4 hold and $\beta = 1$. For any HDLSS-consistent estimator $\hat{\alpha}$ of $-\tau^2/n$, $\mathrm{Angle}\,(v_{\hat{\alpha}}, u_{i,\mathcal{S}}) \xrightarrow{P} \pi/2$ $(i = 1, \ldots, m)$ as $p \to \infty$.*

An immediate consequence of Theorem 3.2 is that the projections of any new data $\mathcal{Y}$ onto $v_{\hat{\alpha}}$ exhibit the asymptotic second data piling; as $p \to \infty$, $P_{v_{\hat{\alpha}}}\mathcal{Y}$ pile on two distinct locations, one for each group. Note that the estimator $\hat{\alpha}$ of $-\tau^2/n$ is typically negative, and the second data piling direction is a projected *negatively-ridged* discriminant direction. Since the scaled eigenvalues of S converge to τ^2/n (shown in Lemma C.2), candidates for $\hat{\alpha}$ include $-p^{-1}\hat{\lambda}_i$ for $i = m+1, \ldots, n-2$. In what follows, we fix

$$\hat{\alpha} = -\frac{\widehat{\tau^2}}{n} = -\frac{1}{n-m-2}\sum_{i=m+1}^{n-2} \frac{\hat{\lambda}_i}{p}. \tag{3.3}$$

If L_1 and L_2 do not coincide, then a separating hyperplane orthogonal to $v_{\hat{\alpha}}$, passing $\bar{X}$, is expected to provide a satisfactory classification of the independent data $\mathcal{Y}$. At the end of this subsection, we show that L_i's do not coincide in the limit, thus perfect classification is possible. The distance between the two parallel affine subspaces is asymptotically equivalent to the distance between the two piles of $P_{v_{\hat{\alpha}}}\mathcal{Y}$. This leads to a natural question: Are there other directions $v \in \mathcal{S}_X$ exhibiting the second data piling phenomenon? We will see that there are infinitely many such directions, but among those $v_{\hat{\alpha}}$ leads to the maximal asymptotic distance between the two piles.

The second data piling on a $v \in \mathcal{S}_X \subset \mathbb{R}^p$ is recognized asymptotically as $p \to \infty$. While $v_{\hat{\alpha}}$ is defined for any p, we wish to consider any sequence of directions, which may not have a simple definition applicable to every p, as a candidate for a second data piling direction. For this, it will be clearer to think of $v \in \mathbb{R}^p$ as the pth element of an infinite sequence $\{v\} = (v^{(1)}, \ldots, v^{(p-1)}, v^{(p)}, v^{(p+1)}, \ldots)$.

We write $\{v\}$ for such a sequence with v indicating the pth element of $\{v\}$. Let $\mathcal{V}$ be the collection of sequences of unit vectors in the sample space $\mathcal{X}$, that is, $\mathcal{V} = \{\{v\} : v \in \mathcal{S}_X, \|v\| = 1 \text{ for all } p\}$. Let Y and Y' be any two independent observations in $\mathcal{Y}$ from the same group. We characterize the collection of all sequences of second data piling direction vectors as

$$\begin{aligned}\mathcal{A} = \{\{v\} \in \mathcal{V} : &\text{for any } Y, Y' \text{ with } \pi(Y) = \pi(Y'), \\ &\frac{1}{\sqrt{p}} v^T(Y - Y') \xrightarrow{P} 0, \text{ as } p \to \infty\}.\end{aligned} \tag{3.4}$$

That is, a sequence $\{v\}$ is a second-data-piling-direction sequence if all observations from the same population are projected to a single point asymptotically. In the two-class discrimination problem, there are at most two points to which the independent data are projected. To better understand the second data piling directions, we use alternative but equivalent definitions of $\mathcal{A}$ to show that any $\{v\} \in \mathcal{A}$ is asymptotically close to a sequence $\{w\}$, where each w is in the direct sum of $\text{span}(v_{\hat{\alpha}})$ and $\text{span}(\{\hat{u}_i\}_{i=m+1}^{n-2})$.

Lemma 3.3. *Suppose Assumptions 1–4 hold and $\beta = 1$. Let $\mathcal{A}'$ be the collection of all sequences $\{v\} \in \mathcal{V}$ such that for any independent $\{Y\}$*

$$\lim_{p\to\infty} \text{Var}(p^{-1/2} v^T[Y - E\{Y \mid \pi(Y) = i\}] \mid \pi(Y) = i) = 0,$$

for both $i = 1, 2$, and $\mathcal{A}'' = \{\{v\} \in \mathcal{V} : v^T u_i \xrightarrow{P} 0,\ i = 1, \ldots, m \text{ as } p \to \infty\}$.

(i) $\mathcal{A} = \mathcal{A}' = \mathcal{A}''$.

(ii) For any given $\{v\} \in \mathcal{A}$, there exists a sequence $\{w\} \in \mathcal{B}$ such that $\|w - v\| \xrightarrow{P} 0$ as $p \to \infty$, where $\mathcal{B} = \{\{w\} \in \mathcal{V} : w \in \text{span}(v_{\hat{\alpha}}) \oplus \text{span}(\{\hat{u}_i\}_{i=m+1}^{n-2})\}$.

It is clear that $\{v_{\hat{\alpha}}\} \in \mathcal{B} \subset \mathcal{A}$ is a second data piling direction. Lemma 3.3(ii) shows that other second data piling directions are given by "lifting" $v_{\hat{\alpha}}$ on $\mathcal{S}$ towards $\text{span}(\{\hat{u}_i\}_{i=m+1}^{n-2})$.

We next show that among the many second-data-piling-direction sequences in $\mathcal{A}$, $\{v_{\hat{\alpha}}\}$ maximizes the limiting distance between the two piles. For $\{w\} \in \mathcal{A}$, let $D(w)$ be the probability limit of $p^{-1/2}|w^T(Y_1 - Y_2)|$ as $p \to \infty$, if it exists for Y_i with $\pi(Y_i) = i$ $(i = 1, 2)$. That is,

$$p^{-1/2}|w^T(Y_1 - Y_2)| \xrightarrow{P} D(w) \text{ as } p \to \infty, \tag{3.5}$$

if the limit exists. The probability limit may not exist for, e.g., an oscillating sequence $\{w\}$, and the limiting distance $D(w)$ may be a random variable.

Theorem 3.4. *Suppose Assumptions 1–4 hold and $\beta = 1$. Then, for any $\{w\} \in \mathcal{A}$ such that $D(w)$ exists,*

$$D(w) \le D(v_{\hat{\alpha}})$$

with probability 1, and the equality holds if and only if $\|w - v_{\hat{\alpha}}\| \xrightarrow{P} 0$ *as* $p \to \infty$. *Moreover, for any independent observation* Y,

$$p^{-1/2} v_{\hat{\alpha}}^T (Y - \bar{X}) \xrightarrow{P} \begin{cases} \gamma \frac{n_2}{n}(1-k^2)\delta^2, & \pi(Y) = 1; \\ -\gamma \frac{n_1}{n}(1-k^2)\delta^2, & \pi(Y) = 2, \end{cases} \tag{3.6}$$

as $p \to \infty$, *where* γ *is a strictly positive random variable depending only on the first* m *principal scores of* $\mathcal{X}$.

Consequently, we may call $\{v_{\hat{\alpha}}\}$ a sequence of *maximal* second-data-piling directions. Since $v_{\hat{\alpha}} \in \mathcal{S}$, the result above also justifies our choice of focusing on v_α in $\mathcal{S}$ rather than w_α in the whole sample space $\mathcal{S}_X$.

Theorem 3.4 also shows that the limiting maximal distance between the two piles is $\gamma(1-k^2)\delta^2$. Since the two lines do not coincide in the limit, a perfect classification occurs. To be specific, let $\boldsymbol{\phi}_\alpha(Y; \mathcal{X})$ be a classification rule defined for a given α:

$$\boldsymbol{\phi}_\alpha(Y; \mathcal{X}) = \begin{cases} 1, & v_\alpha^T \left(Y - \bar{X}\right) \geq 0; \\ 2, & v_\alpha^T \left(Y - \bar{X}\right) < 0. \end{cases} \tag{3.7}$$

Theorem 3.5. *Under the setting of Theorem 3.4,* $\mathbb{P}\{\boldsymbol{\phi}_{\hat{\alpha}}(Y; \mathcal{X}) = \pi(Y)\} \to 1$ *as* $p \to \infty$.

A perfect classification occurs in the limit $p \to \infty$, even if the sample size of $\mathcal{X}$ is kept fixed, and for a negative ridge parameter $\hat{\alpha} < 0$.

3.3. Perfect classification at negative ridge

In this subsection, we show that the perfect classification occurs *only* at the negatively ridged $\boldsymbol{\phi}_\alpha$, i.e., at $\alpha = \hat{\alpha}$. Denote the limits of correct classification rates of $\boldsymbol{\phi}_\alpha$ by

$$\begin{aligned} \mathcal{P}_i(\alpha) &= \lim_{p \to \infty} \mathbb{P}\left\{\boldsymbol{\phi}_\alpha(Y; \mathcal{X}) = i \mid \pi(Y) = i\right\} \;\; (i = 1, 2), \\ \mathcal{P}(\alpha) &= \lim_{p \to \infty} \mathbb{P}\{\boldsymbol{\phi}_\alpha(Y; \mathcal{X}) = \pi(Y)\} = \sum_{i=1}^{2} \pi_i \mathcal{P}_i(\alpha). \end{aligned} \tag{3.8}$$

For simplicity, we assume that the prior $\pi_1 = P\{\pi(Y) = 1\} = 1 - \pi_2$ is equal to n_1/n. These limits (3.8) exist as shown in Lemma 3.6 below. There, we provide an explicit expression of $\mathcal{P}_i(\alpha)$, from which the limiting accuracy can be evaluated.

Recall that any observation $X \in \mathbb{R}^p$ is represented by $X = \mu_i + U\Lambda^{1/2} z^{(p)}$, where the elements of $z^{(p)} = (z_{(1)}, \ldots, z_{(p)})^T$ are the first p elements of an infinite sequence; see Assumption 4. For a given $l = 1, \ldots, m$, the uncentered lth principal score of X, $u_l^T X$, depends only on $z_{(l)}$ for each and every p, and the almost sure limit of $p^{-1/2} u_l^T X$ exists. For $X_{ij} \in \mathcal{X}$, we write the lth limiting principal score as $x_{(l),ij}$, which satisfies $\mathbb{P}(\lim_{p \to \infty} p^{-1/2} u_l^T X_{ij} = x_{(l),ij}) = 1$ for

$l = 1, \dots, m$. Furthermore, we denote $x_{ij} = (x_{(1),ij}, \dots, x_{(m),ij})^T$. Similarly, let $y = (y_{(1)}, \dots, y_{(m)})^T$ which collects the first m limiting principal scores of an independent observation $\{Y\}$. In Lemma 3.6 below, the limiting probabilities $\mathcal{P}_i(\alpha)$ depend on the distribution of

$$\xi_\alpha = \left(n\alpha + \tau^2\right)(y - \bar{x})^T \left(\Omega + \left(n\alpha + \tau^2\right) I_m\right)^{-1} (\bar{x}_1 - \bar{x}_2), \tag{3.9}$$

where $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$, $\bar{x} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} x_{ij}/n$ and $\Omega = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$. Note that Ω is the within-class scatter matrix of first m limiting principal scores, so ξ_α in (3.9) can be considered as a generalized linear classifier.

Lemma 3.6. *Suppose Assumptions 1–4 hold.*

(i) If $0 \le \beta < 1$, then $\mathcal{P}_1(\alpha) = \mathcal{P}_2(\alpha) = 1$ for all $\alpha \in \mathbb{R} \setminus \{-\tau^2/n\}$.
(ii) If $\beta = 1$, then for a given $\alpha \in \mathbb{R}$,

$$\begin{aligned} \mathcal{P}_1(\alpha) &= \mathbb{P}\left(\xi_\alpha + C_1 \ge 0 \mid \pi(Y) = 1\right), \\ \mathcal{P}_2(\alpha) &= \mathbb{P}\left(\xi_\alpha - C_2 < 0 \mid \pi(Y) = 2\right), \end{aligned} \tag{3.10}$$

where $C_i = (1 - n_i/n)(1 - k^2)\delta^2 > 0$, $i = 1, 2$, and ξ_α is defined in (3.9).

Lemma 3.6 shows a sharp distinction between a null model with $\beta \in [0, 1)$ and the m-component model with $\beta = 1$. For $\beta < 1$, the within-group variance of X_{ij} becomes negligible when compared to the magnitude of the true mean difference $\|\mu\|$, thus perfect classification occurs for any reasonable classifier. In particular when $\beta < 1$, for any $\alpha \ne -\tau^2/n$, v_α converges to w_{MDP} as $p \to \infty$. The only exception is $\alpha = -\tau^2/n$, at which $v_{-\tau^2/n}$ becomes orthogonal to w_{MDP} as $p \to \infty$. The result on $\beta < 1$ is consistent with findings in Hall, Marron and Neeman (2005); Jung (2018), where only the case $\beta = 0$ was studied. Note that in those previous work an additional condition $\delta^2 > |1/n_1 - 1/n_2|\tau^2$ is required, while we do not. This is because we use the total mean $\bar{X}$ rather than the centroid mean $(\bar{X}_1 + \bar{X}_2)/2$ used in Hall, Marron and Neeman (2005); Jung (2018).

The m-component model with $\beta = 1$ is more interesting, as the limiting accuracy $\mathcal{P}(\alpha) = \pi_1\mathcal{P}_1(\alpha) + \pi_2\mathcal{P}_2(\alpha)$ depends on the distribution of the first m true principal component scores $(z_{(1)}, \dots, z_{(m)})^T$, through ξ_α. Note that the sign of ξ_α depends on an m-dimensional classification result based on the unobservable principal component scores of Y and $\mathcal{X}$. The positive constant C_i is related to the maximal second-data-piling distance (3.6), and represents the magnitude of group separation on the nullspace of span$(u_1, \dots, u_m)$.

With a regularizing condition on the distribution of $(z_{(1)}, \dots, z_{(m)})$, the limiting accuracy $\mathcal{P}(\alpha)$ is uniquely maximized at $\alpha = -\tau^2/n$. Note that at $\alpha = -\tau^2/n$, $\xi_\alpha = 0$ and $\mathcal{P}(-\tau^2/n) = 1$.

Theorem 3.7. *Suppose Assumptions 1–4 hold and $\beta = 1$. If $\{x : f_z(x) > 0\} = \mathbb{R}^m$, where f_z is the joint density of $(z_{(1)}, \dots, z_{(m)})^T$, then $\alpha = -\tau^2/n$ is the unique maximizer of $\mathcal{P}(\alpha)$.*

TABLE 1

Estimates of the accuracy (standard error) of the three classifiers. As p increases, the accuracy of the classifier based on $v_{\hat{\alpha}}$ becomes the highest, and approaches to 1.

p	w_{MDP}	$w_{\hat{\alpha}}$	$v_{\hat{\alpha}}$
100	0.594 (0.027)	0.517 (0.037)	0.586 (0.028)
500	0.722 (0.021)	0.537 (0.038)	0.710 (0.021)
2000	0.847 (0.035)	0.579 (0.060)	0.871 (0.018)
5000	0.905 (0.048)	0.613 (0.071)	0.965 (0.008)
10000	0.917 (0.061)	0.645 (0.098)	0.994 (0.003)

A wide range of distributions, including the normal, satisfy the condition of Theorem 3.7.

Remark 3.1. *The regularity condition in Theorem 3.7 can be written more directly for ξ_α having its support on $(-\infty, \infty)$ for any $\alpha \neq -\tau^2/n$, and is relaxed by the following condition: For any $\alpha \neq -\tau^2/n$, the density f_{ξ_α} of ξ_α satisfies either $(-\infty, -C_1) \cap \{x : f_{\xi_\alpha}(x \mid \pi(Y) = 1) < 0\} \neq \emptyset$ or $(C_2, \infty) \cap \{x : f_{\xi_\alpha}(x \mid \pi(Y) = 2) > 0\} \neq \emptyset$.*

4. Numerical studies

In this section, we first numerically demonstrate the perfect classification of $\boldsymbol{\phi}_\alpha$ (3.7) via a simulation experiment (Section 4.1), then confirm that the optimal ridge parameter α of the classifier $\boldsymbol{\phi}_\alpha$ is indeed negative in some real data situation, including well-known handwritten image datasets (Section 4.2) and a number of microarray datasets (Section 4.3).

4.1. A simulation experiment

We compare the classification performances of the classifier $\boldsymbol{\phi}_{\hat{\alpha}}$ (3.7) based on $v_{\hat{\alpha}}$, and two others classifiers, defined similar to (3.7) but using $w_{\hat{\alpha}}$ and w_{MDP} in place of $v_{\hat{\alpha}}$, respectively. The model we use is a one-component model satisfying Assumptions 1–4 with $\beta = 1$. Specifically, $X_{ij} \sim \mathcal{N}_p(\mu_i, \Sigma)$, where Σ has a compound symmetry structure, $\Sigma = 1_p 1_p^T + 40 I_p$, with eigenvalues $(p+40, 40, \ldots, 40)$ and the first eigenvector $u_1 = 1_p/\sqrt{p}$. Here, 1_p is the p-vector consisting of 1. The first $p/2$ coordinates of the mean difference vector $\mu = \mu_1 - \mu_2$ are $(\sqrt{2} + \sqrt{3})/2$, and the rest $(\sqrt{2} - \sqrt{3})/2$. That is, $\mu = (\sqrt{p}/2)u_1 + \sqrt{(3p)/4}u_2$, where $u_2^T = (1_{p/2}^T, -1_{p/2}^T)/\sqrt{p}$. We set $n_1 = n_2 = 20$, and p varies from 100 to 10,000. For this model, $\delta^2 = 1$, $k = 1/2$, $\sigma = 1$ and $-\tau^2/n = -1$.

The correct classification rates of the three classifiers, obtained from the sample of size $n = 40$, are empirically computed using 1,000 independent observations. This is averaged over 100 repetitions to estimate the accuracy $\mathbb{P}\{\boldsymbol{\phi}(Y; \mathcal{X}) = \pi(Y)\}$ of the classifier $\boldsymbol{\phi}$. Table 1 displays the result. Using $v_{\hat{\alpha}}$ not only outperforms the others, but also achieves nearly perfect classification for sufficiently large p. The poor performance of using $w_{\hat{\alpha}}$ is due to the inflation of noisy second term in (2.2).

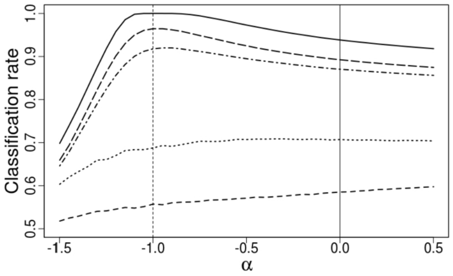


Fig 3. *A graph showing the correct classification rates of the classifier ϕ_α (using v_α) for $p = 100$ (dashed), 500 (dots), 2,000 (dot-dashed), and 5,000 (long dashed). The limiting accuracy $\mathcal{P}(\alpha)$ is shown as a solid curve. The vertical reference lines indicate $\alpha = 0$ (solid) and $\alpha = -\tau^2/n = -1$ (dashed).*

Under the same model, Fig. 3 displays an estimate of the accuracy of ϕ_α at $\alpha \in (-1.5, 0.5)$ for several choices of dimension. At $\alpha = 0$, ϕ_0 corresponds to the classifier based on w_{MDP}. For moderately large dimensions ($p = 200, 500$), the accuracy increases as α increases, and positive ridge-parameters appear to be optimal. On the other hand, for p in the thousands, the correct classification rate is the highest near $\alpha = -\tau^2/n = -1$, as Theorem 3.7 postulates. For this model, near-perfect classification will occur at $\alpha = -\tau^2/n$, for larger choices of p.

4.2. Handwritten character recognition examples

The original MNIST (Modified National Institute of Standards and Technology) dataset (LeCun, Cortes and Burges, 2010) contains 60,000 images of handwritten digits 0 to 9, while the EMNIST (Extended MNIST) database (Cohen et al., 2017) has 124,800 images of handwritten English alphabets a to z, in $28 \times 28 = 784$ pixels. Since the true classification boundaries for these data are likely to be non-linear, we employed random Fourier features (Rahimi and Recht, 2007), a popular method originally developed for kernel methods for large scale problems. The idea of random Fourier features is to embed the original data into high-dimensional space so that a linear inner product in the embedded space approximates a non-linear kernel function. As similarly done in Kobak, Lomond and Sanchez (2020), we obtained random Fourier features of each image corresponding to Gaussian RBF kernel and used them as variables for linear classification. Let x_j be the 1×784 vector with pixel intensity values for the jth image, scaled to be bounded by -1 and 1. For each choice of $d = 500, 1000, 2000, 3000$, we then calculated $\exp(-ix_jZ)$, where $Z \in \mathbb{R}^{784\times d}$ is a random matrix consisting of independent Gaussian random variables with mean zero and standard deviation 0.1. Taking the real and the imaginary parts of $\exp(-ix_jZ)$, we obtained $p = 2d$ features.

For classification, we converted the multi-category problems with 10 groups to $\binom{10}{2} = 45$ binary problems for the MNIST data and 26 groups to $\binom{26}{2} = 325$

Table 2

Number of cases for which the optimal α is negative. The result is based on average test errors from 10 repetitions.

	p	$m = 1$	$m = 2$	$m = 3$	$m = 4$
MNIST (45 cases)	1000	25	30	30	33
	2000	29	37	40	41
	4000	29	37	38	41
	6000	33	36	41	42
EMNIST (325 cases)	1000	195	203	207	195
	2000	241	270	280	289
	4000	253	280	295	307
	6000	252	291	307	311

binary problems for the EMNIST data. To mimic the HDLSS situation, we randomly chose $n_1 = n_2 = 100$ images for each class as the training data for each binary problem, and computed v_α (2.5) and the corresponding classifier $\boldsymbol{\phi}_\alpha$ (3.7), for each of $m = 1, 2, 3, 4$ and for a fine grid of $\alpha \in (-20, 20)$. Using the hold-out data as the test data set, the test error rates are computed for each choice of (m, α). This experiment was repeated ten times to report the average test error rates.

For all choices of dimension p, and for all choice of the number of components m, the optimal choice of the ridge parameter turns out to be negative for a majority of cases considered; Table 2 collects the number of cases under which a negative ridge parameter provides the smallest test error rate. For a reference, Figure 4 shows the test error rates of $\boldsymbol{\phi}_\alpha$ over the grid of α, when the number of leading components is set as $m = 4$. Patterns are similar for $m = 1, 2, 3$; see Figs. D.1—D.3 in Appendix D.1. As the dimension increases, the overall error rates decrease, and the fraction of "negative-ridge-optimal" cases (i.e., the optimal choice of α is negative) increases as the dimension increases.

We also have repeated the above experiment with a larger sample size of 200 for each category, for the MNIST dataset. When the dimension p is large enough, by increasing the sample size from $n_i = 100$ to $n_i = 200$, the optimal ridge parameter $\tilde{\alpha}_n$ tends to shrink to zero; see the top panels of Fig. 5. For most cases, $\tilde{\alpha}_{100} < \tilde{\alpha}_{200} < 0$. This makes sense since the asymptotically optimal choice of α is $-\tau^2/n$ (see Theorem 3.7), which also shrinks to zero as n increases. We also confirm that as the sample size increases (albeit still much smaller than the dimension), the misclassification rate becomes smaller; see the bottom panels of Fig. 5.

An additional experiment suggests that the key assumptions (Assumptions 1 and 2) are in fact satisfied for our binary classification of the MNIST dataset. See Appendix D.2.

4.3. Microarray data examples

Statistical analysis of microarray gene expression data has been a prominent example of the HDLSS situations. We use eight sets of public microarray data sets to examine whether the conclusion of Section 4.2 holds. The microarray

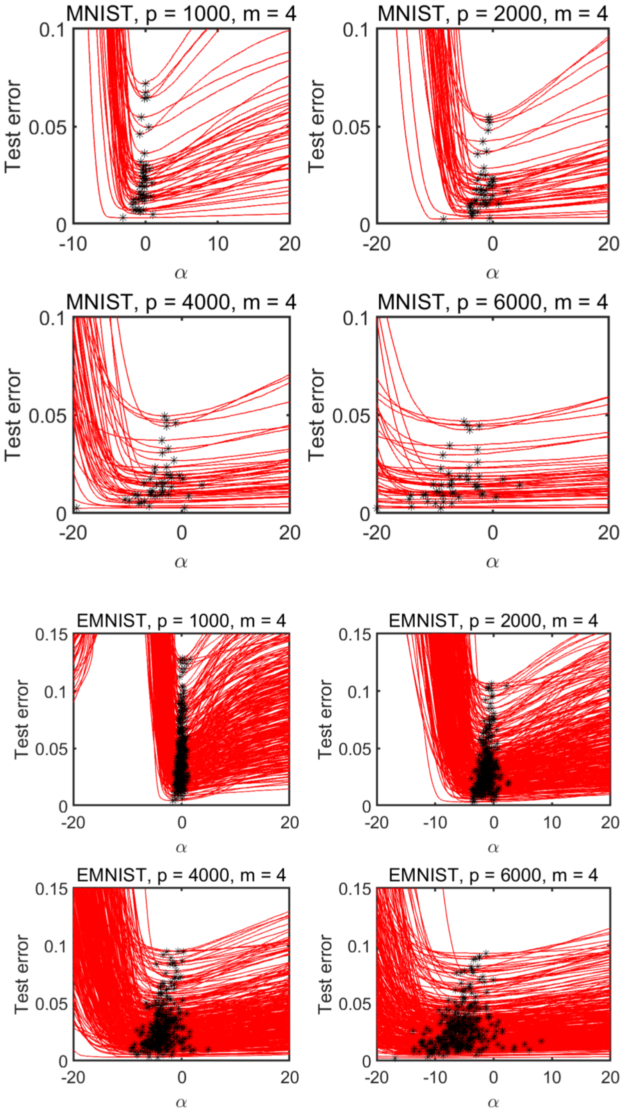


Fig 4. *Misclassification rates of ϕ_α applied to two-group classifications of images of handwritten digits (MNIST) and alphabets (EMNIST). Each curve represents a trajectory of average test error rates from a binary classification. The minimum error rate of each curve is marked with a star.*

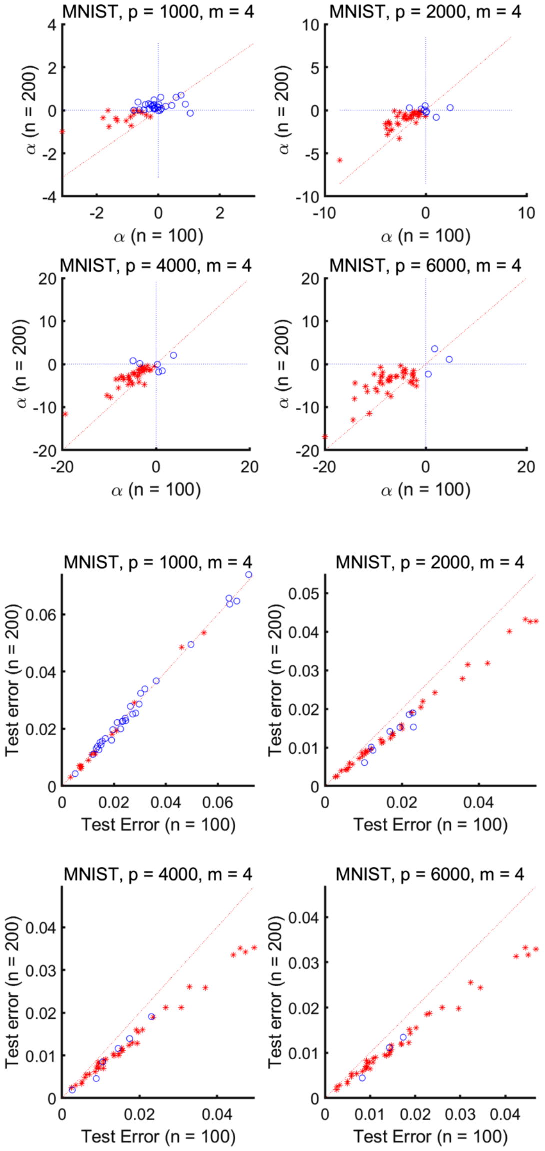


Fig 5. *The effect of sample sizes to the classification. Top panels compare the empirically-optimal ridge parameters $\tilde{\alpha}_n$ with $n = 100$ and 200. Stars represent the cases where both $\tilde{\alpha}_{100}$ and $\tilde{\alpha}_{200}$ are negative; Circles are used for all other cases. Bottom panels compare the misclassification rates of $\phi_{\tilde{\alpha}_n}$ for $n = 100, 200$. Patterns are similar for other choices of $m = 1, 2, 3$.*

TABLE 3
A list of microarray datasets used in our experiment

ID	(p, n, n_1, n_2)	Reference
DLBCL	(2648,77,19,58)	Shipp et al. (2002); Glaab et al. (2012)
prostate	(2135,102,52,50)	Singh et al. (2002)
breast cancer	(47293,128,84,44)	Naderi et al. (2007)
lung cancer	(2530, 37,20,17)	Bhattacharjee et al. (2001)
colon cancer	(2000,62,22,40)	Alon et al. (1999)
methylation	(1413,217,132,85)	Christensen et al. (2009)
breast cancer (Gravier)	(2905,168,111,57)	Gravier et al. (2010)
lymphoma	(7129,77,19,58)	Shipp et al. (2002); Jeffery, Higgins and Culhane (2006)

datasets we used are listed in Table 3 along with their dimensionality and sample sizes.

Since the sample sizes are quite small for these datasets, we computed the leave-one-out average error rates of the classifier (3.7) for a few choice of m and for a fine grid of α. The results are visually summarized in Fig. 6. Two datasets result in zero test error rates for any choice of α and thus excluded from the figure and from further discussion. Out of the remaining six cases, if we take $m = 1$, for all cases the optimal ridge parameters are indeed negative; for $m = 2, 3$, or 4, the number of the negative-ridge-optimal is $5, 5$ or 3, respectively, out of six cases.

The axis for α in Fig. 6 is scaled so that the estimate $\hat{\alpha} = -\hat{\tau}^2/n$ (3.3) is -1. Observe from the figure that the optimal ridge parameters, if they are negative, are typically between $(-\hat{\tau}^2/n, 0)$. This is partly due to the bias of the estimator $-\hat{\tau}^2/n$ (the magnitude of $\hat{\tau}^2/n$ is typically larger than τ^2/n). This suggests that in practice a cross-validation over a range of ridge parameters is recommended, rather than simply using the estimate.

In the data examples above, the true number m of leading principal components is unknown. While the number m may be estimated, by e.g., Bai and Ng (2002); Passemier and Yao (2014); Jung, Lee and Ahn (2018), we do not pursue it here.

5. Discussion

In this work, we have revealed a perhaps counter-intuitive phenomenon of second data piling in the HDLSS context, and showed that a negatively ridged discriminant direction $v_{\hat{\alpha}}$, with $\hat{\alpha} < 0$, exhibits the second data piling with a maximal property. This second data piling direction $v_{\hat{\alpha}}$ is asymptotically orthogonal to the true leading principal component (PC) directions. Thus, by using the direction $v_{\hat{\alpha}}$ for classification, we effectively remove the excessive variability in the leading PC directions. This observation naturally leads to an approach of directly removing the leading PC directions (i.e., projecting the data onto the nullspace of the leading eigenspace). This approach, via an estimation of the PC directions, has been already proposed by Aoshima and Yata (2019) in a similar setting. Although both our classifier and the classifier of Aoshima and Yata

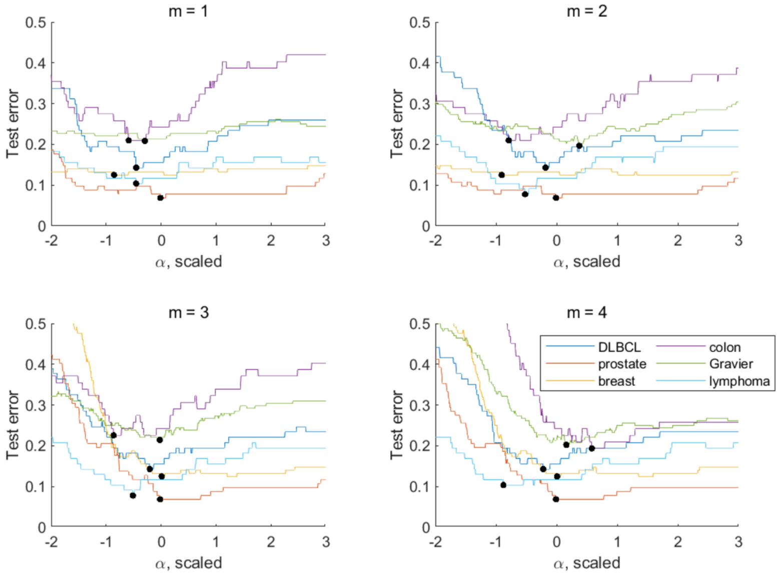


FIG 6. *Misclassification rates of ϕ_α applied to microarray datasets in Table 3. Each curve represents the trajectory of leave-one-out error rates. The minimum error rate of each curve is marked with a filled circle. Results for 'lung cancer' and 'methylation' are omitted as their error rates are zero everywhere.*

(2019) aim to achieve a similar goal of classifying in the nullspace of the leading eigenspace, there has been no discussion of double data piling in Aoshima and Yata (2019). Our current work not only reveals the double data piling phenomenon but also provides an answer to the question "why is removing the leading eigenspace beneficial?"

Our analysis assumes that the true number of components m is known. What happens if one chooses to use an estimate or guess $m^* \neq m$ in defining the subspace $\mathcal{S}$ (2.3) and direction $v_{\hat{\alpha}}$ (2.5)? In fact, the second data piling does not occur for $m^* < m$. Heuristically, if $m^* < m$, then $v_{\hat{\alpha}}$ is not asymptotically orthogonal to the principal component directions with large variance, thus resulting in no data piling. On the other hand, the second data piling occurs for $m^* > m$, with a slower rate of convergence. Rigorously confirming these facts is rather involved, and will be addressed in our subsequent work.

It is well-known that the first data piling also occurs in high-dimensional multi-category data situations. Suppose there are K categories. Let B be the $p \times (K-1)$ matrix with the kth column given by $\bar{X}_k - \bar{X}_K$, and S_T be the total covariance matrix. If $p > n - K - 1$, then the first data piling happens on the column space of $S_T^- B$, where S_T^- is the Moore-Penrose inverse of S_T. Is there a situation under which the second (asymptotic) data piling also occurs for such multi-category classification? While we do have an answer for this question (the answer is yes), we choose to use a binary classification framework to introduce the notion of the second data piling, in this work.

Ahn and Marron (2010) observed that the first maximal data piling direction for $p > n - 2$, when appended by the Fisher's discriminant direction for $p \leq n - 2$, exhibits the so-called double descent phenomenon. The term double descent is coined for the regression setting (Hastie et al., 2019). In the classification context, the double descent phenomenon is understood as follows: The misclassification rate initially decreases as p increases from 1, increases as p approaches to $n - 2$, then decreases as p diverges. The first descent, a "U"-shaped dip, of the misclassification rate is due to the bias-variance trade-off. The second descent occurs at the HDLSS regime. This phenomenon has been observed for a number of real data situations, but has not been fully understood theoretically. A natural question is whether the double descent phenomenon is observed for the second maximal piling direction. Simulation studies (not reported here) using $v_{\hat{\alpha}}$ (well-defined for both $p \leq n$ and $p > n$) indicate that the misclassification rate is monotone with respect to p, thus just a single descent. This observation parallels the empirical observations by Kobak, Lomond and Sanchez (2020) and Wu and Xu (2020) in the regression setting, and suggests that negatively ridged classifiers may mitigate the double decent phenomenon. The misclassification rate corresponding to $v_{\hat{\alpha}}$ is smaller than that using w_{MDP}, for any $p > n$, which is expected for large p by Theorem 3.7. A thorough investigation on this matter is left as a future research agenda.

Appendix A: On ρ-mixing condition

The concept of ρ-mixing was first proposed in Kolmogorov and Rozanov (1960); see Bradley (2005) for detailed introduction on mixing conditions. The definition of ρ-mixing follows. For a σ-algebra $\mathcal{A}$, denote the class of square-integrable and $\mathcal{A}$-measurable functions as $\mathcal{L}^2(\mathcal{A})$. The ρ-type measure of dependency of two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ is defined as

$$\rho\left(\mathcal{F}, \mathcal{G}\right) := \sup\{|\mathrm{Corr}(f, g)| : f \in \mathcal{L}^2(\mathcal{F}), g \in \mathcal{L}^2(\mathcal{G})\}.$$

Suppose $(Y_k, k \in \mathbb{Z})$ is a sequence of random variables. For $-\infty \leq J \leq L \leq \infty$, denote the σ-algebra generated by $\{Y_k : J \leq k \leq L\}$ as $\mathcal{F}_J^L$. Then, the ρ-mixing coefficient is defined as:

$$\rho(n) := \sup_{j \in \mathbb{Z}} \rho(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{\infty}).$$

The sequence of random variables $(Y_k, k \in \mathbb{Z})$ is called ρ-mixing if $\rho(n) \longrightarrow 0$ as $n \to \infty$.

Appendix B: Additional details for the linear discriminant direction $\boldsymbol{w}_{\boldsymbol{\alpha}}$

In this section, we provide more descriptions of the parameterization path w_α has.

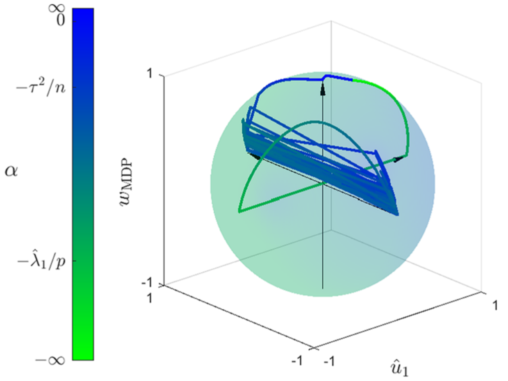


Fig B.1. *The parameterization path of $\alpha \mapsto (w_\alpha^T \hat{u}_1, c_\alpha, w_\alpha^T w_{\mathrm{MDP}})$, where c_α is defined in* (B.1). *Whenever $-\alpha_p$ passes the eigenvalues of S, the direction w_α is reversed, and exactly $n-2$ flips occur.*

B.1. Parametrization path of w_α with sign flips

Main article Figure 1 plots a parametrization path of w_α. For its second coordinate, we use

$$b_\alpha = \left| \sum_{i=2}^{n-2} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d \right| \Bigg/ \left| \sum_{i=1}^{n-2} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d + \hat{U}_2 \hat{U}_2^T d \right|,$$

which is exactly $\|(I - P_{\hat{u}_1} - P_{w_{\mathrm{MDP}}}) w_\alpha\|$. The term b_α is for the second term in (2.2), which lies on an $(n-3)$-dimensional subspace. Suppressing it to a non-negative real value, however, may blind the noticeable features on the parametrization of w_α such as sign-flips. Here, we use a different choice of coordinate c_α,

$$c_\alpha = \Pi_{i=2}^{n-2} \mathrm{sgn}\left(\hat{\lambda}_i + \alpha_p\right) \cdot \left| \sum_{i=2}^{n-2} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d \right| \Bigg/ \left| \sum_{i=1}^{n-2} \frac{\alpha_p}{\hat{\lambda}_i + \alpha_p} \hat{u}_i \hat{u}_i^T d + \hat{U}_2 \hat{U}_2^T d \right|. \tag{B.1}$$

As depicted in Figure B.1, the coordinate c_α is designed so that the sign is reversed every time α_p passes $-\hat{\lambda}_i$, which reflects the property of w_α.

B.2. Projection of w_α into lower dimensional space

Another cost we pay when suppressing the $(n-3)$-dimensional vector, the second term in (2.2), into one coordinate is that even different vectors have the same coordinates. For instance, $b_\alpha \approx 1$ whenever α_p reaches $-\hat{\lambda}_i$ for $i = 2, \dots, n-2$. Figure B.2 illustrates the curve with only three axis $\hat{u}_1$, $\hat{u}_2$, w_{MDP}. A similar shape of the path can be obtained by altering $\hat{u}_2$ to $\hat{u}_i$, $3 \le i \le n-2$.

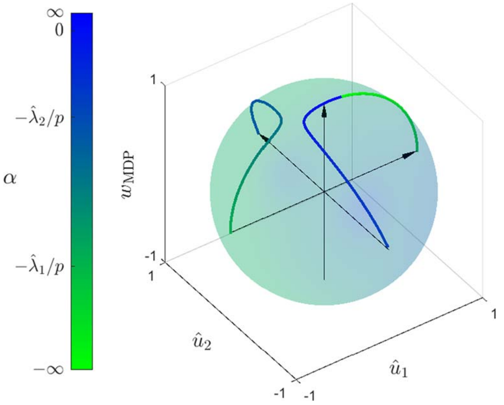


FIG B.2. *The parameterization path of* $\alpha \mapsto (w_\alpha^T \hat{u}_1, w_\alpha^T \hat{u}_2, w_\alpha^T w_{\mathrm{MDP}})$. *Exactly two sign flips are illustrated.*

Appendix C: Technical Details

C.1. Preliminary results on principal component scores

Denote horizontally concatenated $p \times n$ data matrix as X, that is,

$$X = [X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}].$$

The within sample covariance matrix S can be expressed as follows: $S = (X - \widetilde{X})(X - \widetilde{X})^T/n$ where $\widetilde{X} = [\bar{X}_1, \dots, \bar{X}_1, \bar{X}_2, \dots, \bar{X}_2]$. Consider the eigenvalue decomposition of Σ, $\Sigma = U \Lambda U^T$ where $\Lambda = \mathrm{Diag}\{\lambda_i\}_{i=1}^p$, $\lambda_i = p^\beta \sigma_i^2 + \tau_i^2$ $(i = 1, \dots, m)$ and $\lambda_i = \tau_i^2$ $(i = m+1, \dots, p)$. Also, $U = [u_1, \dots, u_p]$ is the $p \times p$ orthogonal matrix where u_i's are normalized eigenvectors corresponding to λ_i. Denote the matrix of principal component scores of X, Z, that is,

$$Z = \Lambda^{-1/2} U^T (X - [\mu_1 1_{n_1}^T, \mu_2 1_{n_2}^T]) = \begin{bmatrix} z_1^T \\ \vdots \\ z_p^T \end{bmatrix} = \begin{bmatrix} z_{1,1}^T & z_{2,1}^T \\ \vdots & \vdots \\ z_{1,p}^T & z_{2,p}^T \end{bmatrix}$$

where $z_{i,j}$ is a vector of principal component scores corresponding to the j-th principal component and the i-th class. Finally, write a vector of principal component scores of new observation Y as $\zeta = (\zeta_1, \dots, \zeta_p)^T$. Then, each element of Z and ζ are uncorrelated and have mean 0 and unit variance.

Lemma C.1. *Under Assumptions 1—4, the following hold simultaneously. The limits are with respect to $p \to \infty$.*

(a) $\frac{1}{p} \mu^T U \Lambda^{1/2} \zeta \xrightarrow{P} \begin{cases} \sum_{i=1}^m \sigma_i \cdot k_i \delta \cdot \zeta_i, & \beta = 1; \\ 0, & 0 \le \beta < 1. \end{cases}$

(b) $\frac{1}{p}\mu^T U \Lambda^{1/2} Z \xrightarrow{P} \begin{cases} \sum_{i=1}^m \sigma_i \cdot k_i \delta \cdot z_i^T, & \beta = 1; \\ 0, & 0 \le \beta < 1. \end{cases}$

(c) $\frac{1}{p} Z^T \Lambda \zeta \xrightarrow{P} \begin{cases} \sum_{i=1}^m \sigma_i^2 z_i \zeta_i, & \beta = 1; \\ 0, & 0 \le \beta < 1. \end{cases}$

(d) $\frac{1}{p} Z^T \Lambda Z \xrightarrow{P} \begin{cases} \sum_{i=1}^m \sigma_i^2 z_i z_i^T + \tau^2 I_n, & \beta = 1; \\ \tau^2 I_n, & 0 \le \beta < 1. \end{cases}$

Proof. We begin by decomposing the quantities in (a)–(d) into two terms:

$$\begin{aligned} \frac{1}{p}\mu^T U \Lambda^{1/2}\zeta &= \frac{1}{p}\sum_{i=1}^m (u_i^T \mu)\sqrt{p^\beta \sigma_i^2 + \tau_i^2}\zeta_i + \frac{1}{p}\sum_{i=m+1}^p \tau_i (u_i^T\mu)\zeta_i := A_1 + A_2, \\ \frac{1}{p}\mu^T U \Lambda^{1/2} Z &= \frac{1}{p}\sum_{i=1}^m (u_i^T \mu)\sqrt{p^\beta \sigma_i^2 + \tau_i^2} z_i^T + \frac{1}{p}\sum_{i=m+1}^p \tau_i (u_i^T\mu) z_i^T := B_1 + B_2, \\ \frac{1}{p} Z^T \Lambda \zeta &= \sum_{i=1}^m \sigma_i^2 p^{\beta-1} z_i \zeta_i + \frac{1}{p}\sum_{i=1}^p \tau_i^2 z_i \zeta_i := C_1 + C_2, \\ \frac{1}{p} Z^T \Lambda Z &= \sum_{i=1}^m \sigma_i^2 p^{\beta-1} z_i z_i^T + \frac{1}{p}\sum_{i=1}^p \tau_i^2 z_i z_i^T := D_1 + D_2. \end{aligned}$$

The terms A_2, B_2, C_2 are irrelevant to β, and can be shown to converge to 0 as $p \to \infty$. Specifically, to handle A_2 and B_2, for any fixed $l \ge 1$ let $W_i \in \mathbb{R}^l$ $(i = 1, \dots, p)$ be any random vectors satisfying $E(W_i) = 0$, $\mathrm{Var}(W_i) = I_l$ and $\mathrm{Cov}(W_i, W_\iota) = 0$ for any $1 \le i \ne \iota \le p$. Note that by Assumption 1, there exists $M < \infty$ such that $\tau_i \le M$ for all i. Chebyshev's inequality gives

$$\begin{aligned} &\mathbb{P}\left(\left\| \frac{1}{p}\sum_{i=m+1}^p \tau_i (u_i^T\mu) W_i \right\| > \epsilon \right) \\ &\le \frac{1}{p^2\epsilon^2} E\left[\left\{ \sum_{i=m+1}^p \tau_i (u_i^T\mu) W_i \right\}^T \left\{ \sum_{i=m+1}^p \tau_i (u_i^T\mu) W_i \right\} \right] \\ &= \frac{1}{p^2\epsilon^2} E\left\{ \sum_{i=m+1}^p \tau_i^2 (u_i^T\mu)^2 W_i^T W_i + \sum_{m+1 \le i \ne \iota \le p} \tau_i \tau_\iota (u_i^T\mu)(u_\iota^T\mu) W_i^T W_\iota \right\} \\ &= \frac{1}{p^2\epsilon^2} E\left\{ \sum_{i=m+1}^p \tau_i^2 (u_i^T\mu)^2 w_i^T w_i \right\} \le \frac{lM^2}{p^2\epsilon^2} \sum_{i=m+1}^p (u_i^T\mu)^2 \le \frac{lM^2}{p\epsilon^2}\frac{1}{p}\|\mu\|^2 \longrightarrow 0, \end{aligned}$$

as $p \to \infty$. Letting $W_i = \zeta_i$ or $W_i = z_i$, we have $A_2 \to 0$ and $B_2 \to 0$ in probability as $p \to \infty$. Similarly, $C_2 \to 0$ in probability as well since

$$\mathbb{P}\left(\left\| \frac{1}{p}\sum_{i=1}^p \tau_i^2 z_i \zeta_i \right\| > \epsilon \right) \le \frac{1}{p^2\epsilon^2} E\left\{ \left(\sum_{i=1}^p \tau_i^2 z_i \zeta_i \right)^T \left(\sum_{i=1}^p \tau_i^2 z_i \zeta_i \right) \right\}$$

$$
\begin{aligned}
&= \frac{1}{p^2\epsilon^2} E\left(\sum_{i=1}^{p} \tau_i^4 z_i^T z_i \zeta_i^2 + \sum_{i \neq \iota} \tau_i^2 \tau_\iota^2 z_i^T z_\iota \zeta_i \zeta_\iota\right) \\
&= \frac{1}{p^2\epsilon^2} E\left(\sum_{i=1}^{p} \tau_i^4 z_i^T z_i \zeta_i^2\right) \le \frac{nM^4}{p\epsilon^2} \longrightarrow 0 \quad \text{as} \quad p \to \infty,
\end{aligned}
$$

in which we used the fact that z_i and ζ_i are independent.

The term D_2 is irrelevant to β as well, but does not degenerate. We use Theorem 1 of Jung and Marron (2009) which states that, in our context, if

$$
\frac{\sum_{i=1}^{p} \tau_i^4}{\left(\sum_{i=1}^{p} \tau_i^2\right)^2} \to 0 \quad \text{as} \quad p \to \infty, \tag{C.1}
$$

and Assumptions 1–4 are satisfied, then $D_2 = \sum_{i=1}^{p} \tau_i^2 z_i z_i^T / p \xrightarrow{P} I_n$ as $p \to \infty$. Since by Assumption 1, $\tau_i \le M$ and $\lim_{p\to\infty} \sum_{i=1}^{p} \tau_i^2/p = \tau^2$, (C.1) holds.

It remains to show that A_1, B_1, C_1, and D_1 converge to their respective counterparts in the statement of the lemma. For $\beta < 1$, they all converge to 0 almost surely. For $\beta = 1$, Assumptions 3–4 guarantee that with probability 1, they converge to the random variables given in the statement. Combining the above gives the desired result. □

Let J_m be the matrix of ones of size $m \times m$. For

$$
J = \begin{bmatrix} \frac{1}{n_1} J_{n_1} & O \\ O & \frac{1}{n_2} J_{n_2} \end{bmatrix}, \tag{C.2}
$$

the group-wise centered data matrix is $X - \widetilde{X} = X(I_n - J)$, where

$$
\widetilde{X} = XJ = [\bar{X}_1, \ldots, \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_2].
$$

The empirical common principal components are given either by the eigendecomposition of S or by the singular-value-decomposition of $X - \widetilde{X} = \hat{U}_1 D \hat{V}_1^T = \sum_{i=1}^{n-2} d_i \hat{u}_i \hat{v}_i^T$, where $\hat{u}_i$ is the ith sample principal component direction and $\hat{v}_i$ is the vector of (normalized) sample principal component scores. Here, the $(n-2) \times (n-2)$ diagonal matrix D collects the non-zero singular values $\{d_i\}_{i=1}^{n-2}$ in descending order. Note that

$$
\begin{aligned}
\hat{u}_i = d_i^{-1}(X - \widetilde{X})\hat{v}_i &= \frac{1}{\sqrt{n}} \hat{\lambda}_i^{-1/2} (X - \widetilde{X}) \hat{v}_i \\
&= \frac{1}{\sqrt{n}} \hat{\lambda}_i^{-1/2} U \Lambda^{1/2} Z (I_n - J) \hat{v}_i,
\end{aligned} \tag{C.3}
$$

where $\hat{\lambda}_i = d_i^2/n$ is the ith largest eigenvalue of S.

We make use of the dual matrix $S_D = (X - \widetilde{X})^T (X - \widetilde{X})/n$ which is a finite-dimensional matrix and shares all $n-2$ nonzero eigenvalues with S. By writing

$X - \widetilde{X} = X(I_n - J)$ and by Lemma C.1 (d), the limiting distribution of S_D can be obtained. As $p \to \infty$,

$$\begin{aligned} \frac{nS_D}{p} &= \frac{1}{p}(I_n - J)Z^T \Lambda Z(I_n - J) \\ &\xrightarrow{P} \begin{cases} \sum_{i=1}^m \sigma_i^2 (I_n - J) z_i z_i^T (I_n - J) + \tau^2 (I_n - J) =: S_0, & \beta = 1; \\ \tau^2 (I_n - J), & 0 \le \beta < 1. \end{cases} \end{aligned} \tag{C.4}$$

For a square matrix M, we denote $\phi_i(M)$ and $v_i(M)$ as the ith largest eigenvalue of M and corresponding eigenvector, respectively. Also, let $v_{ij}(M)$ be the jth coefficient of $v_i(M)$. The following lemmas give the probability limits of eigenvalues and eigenvectors of the sample covariance matrix S, respectively.

Lemma C.2. *Suppose Assumptions 1—4 are satisfied. Let an $n \times m$ matrix of the leading m component scores as $W = [\sigma_1 z_1, \dots, \sigma_m z_m]$ and $\Omega = W^T(I_n - J)W$.*

(i) If $\beta = 1$, then

$$\frac{n\hat{\lambda}_i}{p} \xrightarrow{P} \begin{cases} \phi_i(\Omega) + \tau^2, & i = 1, \dots, m; \\ \tau^2, & m+1 \le i \le n-2, \end{cases}$$

as $p \to \infty$.

(ii) If $0 \le \beta < 1$, then $n\hat{\lambda}_i/p \xrightarrow{P} \tau^2$ as $p \to \infty$ for $i = 1, \dots, n-2$.

Proof. As can be seen in (C.4), we have

$$\begin{aligned} nS_D/p &= (I_n - J)Z^T \Lambda Z (I_n - J)/p \\ &\xrightarrow{P} \begin{cases} (I_n - J)(WW^T + \tau^2 I_n)(I_n - J) = S_0, & \beta = 1 \\ \tau^2 (I_n - J), & 0 \le \beta < 1, \end{cases} \end{aligned} \tag{C.5}$$

as $p \to \infty$. Here, (ii) follows immediately from (C.5). While for $\beta = 1$, we get $p^{-1} n \phi_i(S_D) \xrightarrow{P} \phi_i(S_0)$ for $i = 1, \dots, n-2$ as $p \to \infty$. To show (i), we claim that for $i = 1, \dots, m$, $\phi_i(S_0) = \phi_i(\Omega) + \tau^2$. For this, let λ be a nonzero eigenvalue of $(I_n - J)WW^T(I_n - J)$ and v be its corresponding eigenvector. Then, there exists u satisfying $v = (I_n - J)u$, and thus $(I_n - J)(WW^T)(I_n - J)u = \lambda(I_n - J)u$. Hence,

$$S_0 v = (I_n - J)(WW^T + \tau^2 I_n)(I_n - J)u = (\lambda + \tau^2)(I_n - J)u = (\lambda + \tau^2)v.$$

Since $\Omega = W^T(I_n - J)W$ and $(I_n - J)WW^T(I_n - J)$ share their eigenvalues, we have $\phi_i(S_0) = \phi_i(\Omega) + \tau^2$ for $i = 1, \dots, m$. On the other hand, since Ω is of rank m with probability 1, the rest of eigenvalues of S_0 equals to τ^2. □

For $\beta = 1$, let $\widetilde{V}_1 = [v_1(S_0), \dots, v_{n-2}(S_0)]$ which collects the eigenvectors of S_0 in the proof of Lemma C.2 and $\widetilde{D} = \text{Diag}\{\tilde{d}_i\}_{i=1}^p$ where $\tilde{d}_i = \sqrt{\phi_i(\Omega) + \tau^2}$

for $i = 1, \dots, m$ and $\tilde{d}_i = \tau$ for $i \geq m+1$. Then, we have the limits of right singular vectors and singular values,

$$\hat{V}_1 \xrightarrow{P} \widetilde{V}_1 \quad \text{and} \quad \frac{1}{\sqrt{p}} D \xrightarrow{P} \widetilde{D}, \tag{C.6}$$

as $p \to \infty$. The limit of eigenvector $\hat{u}_i$ is analyzed through $\hat{v}_i$ utilizing (C.3) and (C.6).

Finally, the following lemma suggests the limits of the sample eigenvectors $\hat{u}_i$ and the inner product $d^T \hat{u}_i$ for $\beta \in [0, 1]$. The results in Lemma C.3 will be frequently used in the later sections.

Lemma C.3. *Let Assumptions 1—4 hold.*

(i) The limit of the inner product between the population eigenvector u_ι ($\iota = 1, \dots, m$) and the sample eigenvector $\hat{u}_i$ ($i = 1, \dots, m$) depends on β;

$$u_\iota^T \hat{u}_i \xrightarrow{P} \begin{cases} \sqrt{\phi_i(\Omega)/(\phi_i(\Omega) + \tau^2)}\, v_{i\iota}(\Omega), & \beta = 1; \\ 0, & 0 \leq \beta < 1, \end{cases}$$

as $p \to \infty$. While $u_i^T \hat{u}_\iota \xrightarrow{P} 0$ as $p \to \infty$ for $i \geq m+1$ and $\iota = 1, \dots, n-2$.

(ii) The limit of the inner product between the sample eigenvector $\hat{u}_i$ ($i = 1, \dots, m$) and d depends on β;

$$\frac{1}{\sqrt{p}} d^T \hat{u}_i \xrightarrow{P} \sum_{j=1}^{m} \sqrt{\phi_i(\Omega)/(\phi_i(\Omega) + \tau^2)}\, v_{ij}(\Omega) \{k_j \delta + \sigma_j (\bar{z}_{1,j} - \bar{z}_{2,j})\},$$

for $\beta = 1$, and

$$\frac{1}{\sqrt{p}} d^T \hat{u}_i \xrightarrow{P} 0 \text{ for } 0 \leq \beta < 1 \text{ as } p \to \infty,$$

while $d^T \hat{u}_i / \sqrt{p} \xrightarrow{P} 0$ as $p \to \infty$ for $i \geq m+1$ and $0 \leq \beta \leq 1$.

Proof.

(i) We express $\hat{u}_i$ with $\hat{v}_i$ using (C.3),

$$u_i^T \hat{u}_\iota = \left(\frac{n \hat{\lambda}_\iota}{p} \right)^{-\frac{1}{2}} \frac{1}{\sqrt{p}} u_i^T U \Lambda^{\frac{1}{2}} Z (I_n - J) \hat{v}_\iota = \left(\frac{n \hat{\lambda}_\iota}{p} \right)^{-\frac{1}{2}} \left(\frac{\lambda_i}{p} \right)^{\frac{1}{2}} z_i^T (I_n - J) \hat{v}_\iota.$$

The random variables $n\hat{\lambda}_\iota / p$ and $z_i^T (I_n - J) \hat{v}_\iota$ are stochastically bounded since they converge in probability as shown in Lemma C.2 and (C.6), respectively. For $0 \leq \beta < 1$, since $\lambda_i / p \to 0$ for all $1 \leq i \leq p$, we get $u_i^T \hat{u}_\iota \xrightarrow{P} 0$ for all i. Similarly, $u_i^T \hat{u}_\iota \xrightarrow{P} 0$ for $\beta = 1$ and $m+1 \leq i \leq p$. Now, it suffices to consider the case of $\beta = 1$ and $1 \leq i \leq m$. Due to the duality, $v_i(\Omega)$, $v_i(S_0)$, and $\sqrt{\phi_i(\Omega)}$ are the ith right singular vector, left singular vector, and singular value of $(I_n - J)W$, respectively. Therefore, we have $(I_n - J)Wv_i(\Omega) = \phi_i(\Omega)^{1/2} v_i(S_0)$ for $i = 1, \ldots, m$. Hence, for $i, \iota = 1, \ldots, m$,

$$u_\iota' \hat{u}_i = \phi_i(S)^{-\frac{1}{2}} \left(\frac{\lambda_\iota}{p}\right)^{\frac{1}{2}} z_\iota^T (I_n - J) v_i(S_D) \xrightarrow{P} \frac{\sigma_\iota z_\iota^T (I_n - J)}{\sqrt{\phi_i(\Omega) + \tau^2}} \frac{(I_n - J)W}{\sqrt{\phi_i(\Omega)}} v_i(\Omega)$$

$$= \frac{e_\iota^T \Omega v_i(\Omega)}{\sqrt{\phi_i(\Omega)(\phi_i(\Omega) + \tau^2)}} = \frac{\phi_i(\Omega) e_\iota^T v_i(\Omega)}{\sqrt{\phi_i(\Omega)(\phi_i(\Omega) + \tau^2)}} = \sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega) + \tau^2}} v_{i\iota}(\Omega),$$

as $p \to \infty$.

(ii) Note that $d = \mu + U\Lambda^{1/2} Z \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix}$. We decompose the quantity $d^T \hat{U}_1 / \sqrt{p}$ into two terms:

$$\frac{1}{\sqrt{p}} d^T \hat{u}_i = \frac{1}{\sqrt{p}} \mu^T \hat{u}_i + \frac{1}{\sqrt{p}} \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix}^T Z^T \Lambda^{1/2} U^T \hat{u}_i.$$

From (C.3), (C.6), and Lemmas C.1 and C.2, we have

$$\begin{aligned} \frac{1}{\sqrt{p}} \mu^T \hat{u}_i &= \frac{1}{p} \mu^T U \Lambda^{1/2} Z (I_n - J) \hat{v}_i \sqrt{p} d_i^{-1} \\ &\xrightarrow{P} \begin{cases} \sum_{j=1}^m k_j \delta \sqrt{\phi_i(\Omega)/(\phi_i(\Omega) + \tau^2)} v_{ij}(\Omega) & \beta = 1 \text{ and } i = 1, \ldots, m; \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{C.7}$$

and

$$\begin{aligned} &\frac{1}{\sqrt{p}} \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix}^T Z^T \Lambda^{1/2} U^T \hat{u}_i = \frac{1}{p} \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix}^T Z^T \Lambda Z (I_n - J) \hat{v}_i \sqrt{p} d_i^{-1} \\ &\xrightarrow{P} \begin{cases} \sum_{j=1}^m \sigma_j (\bar{z}_{1,j} - \bar{z}_{2,j}) \sqrt{\phi_i(\Omega)/(\phi_i(\Omega) + \tau^2)} v_{ij}(\Omega), & \beta = 1 \text{ and } i = 1, \ldots, m; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{C.8}$$

Adding (C.7) and (C.8), we get the desired results. □

C.2. Proof of Theorem 2.1

Proof. Let $\beta = 1$. First, we make use of $\tilde{w}_\alpha$ in (2.2). The limiting angle between w_α and $\mathcal{S}$ is analyzed through the quantity,

$$\frac{\tilde{w}_\alpha^T P_{\mathcal{S}} \tilde{w}_\alpha}{\|\tilde{w}_\alpha\| \|P_{\mathcal{S}} \tilde{w}_\alpha\|} = \frac{\|P_{\mathcal{S}} \tilde{w}_\alpha\|}{\|\tilde{w}_\alpha\|}, \tag{C.9}$$

where $P_{\mathcal{S}}$ is the orthogonal projection operator onto $\mathcal{S}$. We claim that the quantity in (C.9) converges to 1 in probability. First, the following holds by Pythagoras' theorem:

$$\frac{1}{p} \|\tilde{w}_\alpha\|^2 = \frac{1}{p} \sum_{i=1}^{n-2} \left(\frac{\alpha}{\hat{\lambda}_i / p + \alpha} \right)^2 (\hat{u}_i^T d)^2 + \frac{1}{p} \left\| \hat{U}_2 \hat{U}_2^T d \right\|$$

$$= \frac{1}{p}\|P_{\mathcal{S}}\tilde{w}_\alpha\|^2 + \frac{1}{p}\sum_{i=m+1}^{n-2}\left(\frac{\alpha}{\hat{\lambda}_i/p+\alpha}\right)^2(\hat{u}_i^T d)^2. \tag{C.10}$$

Since Lemmas C.2 and C.3 give that the last term in (C.10) converges to 0 in probability for $\alpha \neq -\tau^2/n$, it remains to show that $\|P_{\mathcal{S}}\tilde{w}_\alpha\|^2/p$ is stochastically bounded. We decompose $\|P_{\mathcal{S}}\tilde{w}_\alpha\|^2/p$ into two terms:

$$\frac{1}{p}\|P_{\mathcal{S}}\tilde{w}_\alpha\|^2 = \frac{1}{p}\sum_{i=1}^{m}\left(\frac{\alpha}{\hat{\lambda}_i/p+\alpha}\right)^2(\hat{u}_i^T d)^2 + \frac{1}{p}\|\hat{U}_2\hat{U}_2^T d\|^2.$$

The first term converges to a certain quantity from Lemmas C.2 and C.3, so is thus stochastically bounded. To deal with the second term, note that $\|\hat{U}_2\hat{U}_2^T d\|^2 = \|d\|^2 - \|\hat{U}_1\hat{U}_1^T d\|^2$. The limit of norm of d can be derived using Lemma C.1,

$$\begin{aligned}\frac{1}{p}\|d\|^2 &= \frac{1}{p}\|\mu\|^2 + \frac{2}{p}\mu^T U\Lambda^{1/2}Z\begin{bmatrix}\frac{1}{n_1}1_{n_1}\\ -\frac{1}{n_2}1_{n_2}\end{bmatrix} + \frac{1}{p}\begin{bmatrix}\frac{1}{n_1}1_{n_1}\\ -\frac{1}{n_2}1_{n_2}\end{bmatrix}^T Z^T\Lambda Z\begin{bmatrix}\frac{1}{n_1}1_{n_1}\\ -\frac{1}{n_2}1_{n_2}\end{bmatrix}\\ &\xrightarrow{P} \delta^2 + 2\sum_{i=1}^{m}\sigma_i\cdot k_i\delta(\bar{z}_{1,i}-\bar{z}_{2,i}) + \sum_{i=1}^{m}\sigma_i^2(\bar{z}_{1,i}-\bar{z}_{2,i})^2 + \tau^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right),\end{aligned} \tag{C.11}$$

as $p\to\infty$. While we handle the limit of $\|\hat{U}_1\hat{U}_1^T d\|^2$ with Lemma C.3 (ii),

$$\begin{aligned}\frac{1}{p}\|\hat{U}_1\hat{U}_1^T d\|^2 &= \frac{1}{p}\sum_{i=1}^{m}|\hat{u}_i^T d|^2 + o_P(1)\\ &\xrightarrow{P} \sum_{i=1}^{m}\left[\sum_{j=1}^{m}\{k_j\delta + \sigma_j(\bar{z}_{1,j}-\bar{z}_{2,j})\}v_{ij}(\Omega)\right]^2\frac{\phi_i(\Omega)}{\phi_i(\Omega)+\tau^2},\end{aligned} \tag{C.12}$$

as $p\to\infty$. Note that $\sum_{i=1}^m v_{ij}(\Omega)^2 = 1$ and $\sum_{i=1} v_{ij}(\Omega)v_{ij'}(\Omega) = 0$ for $j \neq j'$. Combining this with (C.11) and (C.12) gives

$$\begin{aligned}\frac{1}{p}\|\hat{U}_2\hat{U}_2 d\|^2 &\xrightarrow{P} (1-k^2)\delta^2 + \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\tau^2\\ &+ \sum_{i=1}^{m}\left[\sum_{j=1}^{m}\{k_j\delta + \sigma_j(\bar{z}_{1,j}-\bar{z}_{2,j})\}v_{ij}(\Omega)\right]^2\frac{\tau^2}{\phi_i(\Omega)+\tau^2},\end{aligned} \tag{C.13}$$

as $p\to\infty$. We denote the limit of (C.13) as κ^2 for $\kappa > 0$. Since $\|P_{\mathcal{S}}\tilde{w}_\alpha\|^2/p$ converges in probability, and we come to conclusion. We make use of random variable κ in the proof of Theorem 3.1. □

C.3. Proof of Theorem 3.1

Proof. We continue to assume $\beta = 1$. Note that $u_{\iota,\mathcal{S}} = P_{\mathcal{S}}u_\iota = \sum_{i=1}^m(u_\iota^T\hat{u}_i)\hat{u}_i + (u_\iota^T w_{\mathrm{MDP}})w_{\mathrm{MDP}}$. We begin with focusing on the inner products $u_\iota^T\hat{u}_i$ and $u_\iota^T w_{\mathrm{MDP}}$.

On $u_\iota^T \hat{u}_i$. The limit of the inner product $u_\iota^T \hat{u}_i$ ($\iota, i = 1, \dots, m$) can be derived immediately from Lemma C.3 (i): As $p \to \infty$,

$$u_\iota^T \hat{u}_i \xrightarrow{P} \sqrt{\phi_i(\Omega)/(\phi_i(\Omega) + \tau^2)}\, v_{i\iota}(\Omega), \tag{C.14}$$

where Ω is defined in Lemma C.2.

On $u_\iota^T w_{\mathrm{MDP}}$. We claim that as $p \to \infty$,

$$u_\iota^T w_{\mathrm{MDP}} \xrightarrow{P} \frac{1}{\kappa} \sum_{j=1}^{m} \sum_{l=1}^{m} \{k_l \delta + \sigma_l(\bar{z}_{1,l} - \bar{z}_{2,l})\} v_{j\iota}(\Omega) v_{jl}(\Omega) \frac{\tau^2}{\phi_j(\Omega) + \tau^2}, \tag{C.15}$$

for $\iota = 1, \dots, m$ where κ is defined in (C.13). To show (C.15), recall that $w_{\mathrm{MDP}} = \hat{U}_2 \hat{U}_2^T d / \|\hat{U}_2 \hat{U}_2^T d\|$ and $\hat{U}_2 \hat{U}_2^T d = d - \hat{U}_1 \hat{U}_1^T d$. Since the limits of quantities $u_\iota^T \hat{U}_1$, $\hat{U}_1^T d/\sqrt{p}$ and $\|\hat{U}_2 \hat{U}_2^T d\| \sqrt{p}$ are already analyzed in Lemma C.3 and (C.13), it suffices to show that $u_\iota^T d/\sqrt{p} \xrightarrow{P} k_\iota \delta + \sigma_\iota(\bar{z}_{1,\iota} - \bar{z}_{2,\iota})$, which is given by Assumptions 1—3;

$$\begin{aligned}
\frac{1}{\sqrt{p}} u_\iota^T d &= \frac{1}{\sqrt{p}} u_\iota^T \mu + \frac{1}{\sqrt{p}} u_\iota^T U \Lambda^{1/2} Z \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix} \\
&= \frac{1}{\sqrt{p}} u_\iota^T \mu + \frac{1}{\sqrt{p}} \lambda_\iota^{1/2} z_\iota^T \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix} \xrightarrow{P} k_\iota \delta + \sigma_\iota(\bar{z}_{1,\iota} - \bar{z}_{2,\iota}),
\end{aligned}$$

as $p \to \infty$. We now deal with L_1 and L_2 in (3.1) which generally exist on $\mathcal{S}$,

$$L_i = \{U_{m,\mathcal{S}} t + \kappa_i w_{\mathrm{MDP}} + P_{\mathcal{S}} \bar{X}/\sqrt{p} : t \in \mathbb{R}^m\} \quad (i = 1, 2),$$

where $\kappa_1 = (1 - \eta_1)(1 - k^2)\delta^2/\kappa_{\mathrm{MDP}}$ and $\kappa_2 = -(1 - \eta_2)(1 - k^2)\delta^2/\kappa_{\mathrm{MDP}}$ having $\eta_1 = n_1/n$ and $\eta_2 = n_2/n$. The distance between two piles induced by w_{MDP}, $p^{1/2}\kappa_{\mathrm{MDP}}$, is exactly $\|\hat{U}_2 \hat{U}_2^T d\|$ since $w_{\mathrm{MDP}} = \hat{U}_2 \hat{U}_2^T d / \|\hat{U}_2 \hat{U}_2^T d\|$. Also, as shown in (C.13), $\kappa_{\mathrm{MDP}} \xrightarrow{P} \kappa$ as $p \to \infty$.

For $Y \in \mathcal{Y}$, we temporarily assume $\pi(Y) = 1$. Let $t^o = (t_1, \dots, t_m)^T$ with $t_j = \eta_2 k_j \delta + \sigma_j(\zeta_j - \bar{z}_j)$ and let $v^o = U_{m,\mathcal{S}} t^o + \kappa_1 w_{\mathrm{MDP}} + P_{\mathcal{S}} \bar{X}/\sqrt{p} \in L_1$. We claim that $\|Y_{\mathcal{S}} - v^o\|_2 \xrightarrow{P} 0$ as $p \to \infty$. Since $Y_{\mathcal{S}} - v^o \in \mathcal{S}$, it suffices to show that as $p \to \infty$, (a) $\hat{u}_i^T (Y_{\mathcal{S}} - v^o) \xrightarrow{P} 0$ for $i = 1, \dots, m$ and (b) $w_{\mathrm{MDP}}^T (Y_{\mathcal{S}} - v^o) \xrightarrow{P} 0$.

Note that $\hat{u}_i^T (Y_{\mathcal{S}} - v^o) = p^{-1/2} \hat{u}_i^T (Y - \bar{X}) - \hat{u}_i^T \sum_{j=1}^m t_j u_{j,\mathcal{S}}$ and (C.14) gives the limit of the second term,

$$\hat{u}_i^T \sum_{j=1}^{m} t_j u_{j,\mathcal{S}} = \sum_{j=1}^{m} t_j \hat{u}_i^T u_j \xrightarrow{P} \sum_{j=1}^{m} t_j v_{ij}(\Omega) \sqrt{\phi_i(\Omega)/(\phi_i(\Omega) + \tau^2)}.$$

Also, from Lemma C.1 and (C.7),

$$\begin{aligned}
\frac{1}{\sqrt{p}}\hat{u}_i^T\left(Y-\bar{X}\right) &= \frac{\eta_2}{\sqrt{p}}\hat{u}_i^T\mu + \frac{1}{\sqrt{p}}\hat{u}_i^T U\Lambda^{1/2}\left(\zeta - \frac{1}{n}Z1_n\right) \\
&= \frac{\eta_2}{\sqrt{p}}\hat{u}_i^T\mu + \sum_{j=1}^m \hat{u}_i^T u_j \cdot \sigma_j(\zeta_j - \bar{z}_j) + o_P(1) \\
&= \sum_{j=1}^m \eta_2 k_j \delta v_{ij}(\Omega)\sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega)+\tau^2}} + \sum_{j=1}^m \sigma_j(\zeta_j - \bar{z}_j)v_{ij}(\Omega)\sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega)+\tau^2}} + o_P(1) \\
&\xrightarrow{P} \sum_{j=1}^m \{\eta_2 k_j\delta + \sigma_j(\zeta_j - \bar{z}_j)\}v_{ij}(\Omega)\sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega)+\tau^2}} = \sum_{j=1}^m t_j v_{ij}(\Omega)\sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega)+\tau^2}},
\end{aligned} \tag{C.16}$$

as $p \to \infty$. Therefore, (a) follows.

For (b), note that $w_{\text{MDP}}^T(Y_{\mathcal{S}} - v^o) = p^{-1/2}w_{\text{MDP}}^T(Y - \bar{X}) - w_{\text{MDP}}^T\sum_{j=1}^m t_j u_{j,\mathcal{S}} - \kappa_1$. From (C.15),

$$\begin{aligned}
& w_{\text{MDP}}^T \sum_{j=1}^m t_j u_{j,\mathcal{S}} = \sum_{j=1}^m t_j w_{\text{MDP}}^T u_j \\
& \xrightarrow{P} \frac{1}{\kappa}\sum_{j=1}^m t_j \sum_{l=1}^m \sum_{l'=1}^m \{k_l\delta + \sigma_l(\bar{z}_{1,l} - \bar{z}_{2,l})\}v_{l'j}(\Omega)v_{l'l}(\Omega)\frac{\tau^2}{\phi_{l'}(\Omega)+\tau^2}.
\end{aligned} \tag{C.17}$$

To evaluate the limit of $p^{-1/2}w_{\text{MDP}}^T(Y - \bar{X})$, we decompose it into two terms,

$$p^{-1/2}w_{\text{MDP}}^T(Y-\bar{X}) = p^{-1/2}\frac{(\hat{U}_2\hat{U}_2^T d)^T}{\|\hat{U}_2\hat{U}_2^T d\|}(Y - \bar{X}) = \kappa_{\text{MDP}}^{-1}\left(K_1 - K_2\right),$$

where $K_1 = d^T(Y - \bar{X})/p$ and $K_2 = (\hat{U}_1\hat{U}_1^T d)^T(Y - \bar{X})/p$. Routinely, the limit of K_1 can be evaluated through the expression of d and $Y - \bar{X}$ using the principal scores and Lemma C.1.

$$\begin{aligned}
K_1 &= \frac{1}{p}\left(\mu + U\Lambda^{1/2}Z\begin{bmatrix}\frac{1}{n_1}1_{n_1} \\ -\frac{1}{n_2}1_{n_2}\end{bmatrix}\right)^T\left\{\eta_2\mu + U\Lambda^{1/2}\left(\zeta - \frac{1}{n}Z1_n\right)\right\} \\
&= \frac{1}{p}\Bigg\{\eta_2\|\mu\|^2 + \eta_2\mu^T U\Lambda^{1/2}Z\begin{bmatrix}\frac{1}{n_1}1_{n_1} \\ -\frac{1}{n_2}1_{n_2}\end{bmatrix} + \mu^T U\Lambda^{1/2}\left(\zeta - \frac{1}{n}Z1_n\right) \\
&\quad + \begin{bmatrix}\frac{1}{n_1}1_{n_1} \\ -\frac{1}{n_2}1_{n_2}\end{bmatrix}^T Z^T\Lambda\left(\zeta - \frac{1}{n}Z1_n\right)\Bigg\} \\
&\xrightarrow{P} \eta_2(1-k^2)\delta^2 + \sum_{j=1}^m \{k_j\delta + \sigma_j(\bar{z}_{1,j} - \bar{z}_{2,j})\}\{\eta_2 k_j\delta + \sigma_j\left(\zeta_j - \bar{z}_j\right)\},
\end{aligned} \tag{C.18}$$

as $p \to \infty$. Meanwhile, (C.16) and Lemma C.3 (ii) gives

$$K_2 \xrightarrow{P} \sum_{j=1}^{m} \sum_{l=1}^{m} \sum_{l'=1}^{m} \Bigg[\left\{ k_l \delta + \sigma_l \left(\bar{z}_{1,l} - \bar{z}_{2,l} \right) \right\} \left\{ \eta_2 k_{l'} \delta + \sigma_{l'} \left(\zeta_{l'} - \bar{z}_{l'} \right) \right\} \\ \times \frac{\phi_j(\Omega)}{\phi_j(\Omega) + \tau^2} v_{jl}(\Omega) v_{jl'}(\Omega) \Bigg]. \tag{C.19}$$

Combining (C.13), (C.18) and (C.19), we have

$$p^{-1/2} w_{\mathrm{MDP}}^T (Y - \bar{X}) \xrightarrow{P} \kappa^{-1} \Bigg[\eta_2 (1 - k^2) \delta^2 + \\ \sum_{j=1}^{m} \sum_{l=1}^{m} \sum_{l'=1}^{m} \left\{ k_l \delta + \sigma_l \left(\bar{z}_{1,l} - \bar{z}_{2,l} \right) \right\} \left\{ \eta_2 k_{l'} \delta + \sigma_{l'} \left(\zeta_{l'} - \bar{z}_{l'} \right) \right\} \frac{\tau^2 v_{jl}(\Omega) v_{jl'}(\Omega)}{\phi_j(\Omega) + \tau^2} \Bigg]. \tag{C.20}$$

Simply applying the results in (C.14), (C.15), (C.17) and (C.20), we get (b). The same goes for $Y \in \mathcal{Y}$ with $\pi(Y) = 2$. □

C.4. Proof of Theorem 3.2

Proof. We begin by introducing

$$\tilde{v}_{\alpha} = \sum_{i=1}^{m} \frac{\alpha}{\hat{\lambda}_i / p + \alpha} \left(\frac{1}{\sqrt{p}} \hat{u}_i^T d \right) \hat{u}_i + \frac{1}{\sqrt{p}} \| \hat{U}_2 \hat{U}_2^T d \| w_{\mathrm{MDP}}, \tag{C.21}$$

where $v_{\alpha} \propto \tilde{v}_{\alpha}$. Let $\hat{\alpha}$ be an HDLSS-consistent estimator for $-\tau^2/n$. The quantity of interest is the angle between $\tilde{v}_{\hat{\alpha}}$ and $u_{\iota,S}$,

$$\mathrm{Angle}(\tilde{v}_{\hat{\alpha}}, u_{\iota,S}) = \cos^{-1} \left(\frac{\tilde{v}_{\hat{\alpha}}^T u_{\iota,S}}{\| \tilde{v}_{\hat{\alpha}} \| \| u_{\iota,S} \|} \right), \tag{C.22}$$

for $\iota = 1, \ldots, m$. Combining (C.21) and that $u_{\iota,\mathcal{S}} = \sum_{i=1}^{m} (u_{\iota}^T \hat{u}_i) \hat{u}_i + (u_{\iota}^T w_{\mathrm{MDP}}) w_{\mathrm{MDP}}$, the limit of the inner product $\tilde{v}_{\hat{\alpha}}^T u_{\iota,S}$ becomes

$$\tilde{v}_{\hat{\alpha}}^T u_{\iota,S} = \sum_{i=1}^{m} \frac{\hat{\alpha}}{\hat{\lambda}_i / p + \hat{\alpha}} \left(\frac{1}{\sqrt{p}} \hat{u}_i^T d \right) \left(u_{\iota}^T \hat{u}_i \right) + \frac{1}{\sqrt{p}} \| \hat{U}_2 \hat{U}_2^T d \| \left(u_{\iota}^T w_{\mathrm{MDP}} \right). \tag{C.23}$$

The limit of the right-hand-side of (C.23) can be obtained through (C.13), (C.14), (C.15), Lemmas C.2 and C.3 (ii),

$$\tilde{v}_{\hat{\alpha}}^T u_{\iota,S} \xrightarrow{P} \\ \sum_{i=1}^{m} \frac{-\tau^2}{\phi_i(\Omega)} \sum_{j=1}^{m} \left\{ k_j \delta + \sigma_j \left(\bar{z}_{1,j} - \bar{z}_{2,j} \right) \right\} v_{ij}(\Omega) \sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega) + \tau^2}} v_{i\iota}(\Omega) \sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega) + \tau^2}}$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{m}\left\{k_j\delta+\sigma_j\left(\bar{z}_{1,j}-\bar{z}_{2,j}\right)\right\}v_{ij}(\Omega)v_{i\iota}(\Omega)\frac{\tau^2}{\phi_i(\Omega)+\tau^2}=0,$$

as $p\to\infty$. In order to show that the quantity (C.22) converges to $\pi/2$ in probability, it remains to verify that the denominator in (C.22) does not degenerate as $p\to\infty$. First, (C.14) and (C.15) give that

$$\|u_{\iota,\mathcal{S}}\|^2\xrightarrow{P}\sum_{i=1}^{m}v_{i\iota}(\Omega)^2\frac{\phi_i(\Omega)}{\phi_i(\Omega)+\tau^2}+ \tag{C.24}$$

$$\frac{1}{\kappa^2}\left[\sum_{j=1}^{m}\sum_{l=1}^{m}\{k_l\delta+\sigma_l(\bar{z}_{1,l}-\bar{z}_{2,l})\}v_{j\iota}(\Omega)v_{jl}(\Omega)\frac{\tau^2}{\phi_j(\Omega)+\tau^2}\right]^2>0. \tag{C.25}$$

Also, from (C.13), Lemmas C.2 and C.3 (ii),

$$\begin{aligned}\|\tilde{v}_{\hat{\alpha}}\|^2\xrightarrow{P}&\sum_{i=1}^{m}\frac{\tau^4}{\phi_i(\Omega)(\phi_i(\Omega)+\tau^2)}\left[\sum_{j=1}^{m}\left\{k_j\delta+\sigma_j\left(\bar{z}_{1,j}-\bar{z}_{2,j}\right)\right\}v_{ij}(\Omega)\right]^2+\kappa^2\\ &=:\frac{1}{\gamma^2}>0.\end{aligned} \tag{C.26}$$

By (C.24) and (C.26), the desired result is obtained. The positive term γ depends on the first true principal scores of training data, which are invariant to p. We make use of γ in Theorem 3.5. □

C.5. Proof of Lemma 3.3

Proof. Recall that we assume $\beta=1$.

(i) For Y, $Y'\in\mathcal{Y}$ with $\pi(Y)=\pi(Y')$, denote ζ and ζ' as the vectors consisting of principal scores of Y and Y', respectively. That is, $Y-E(Y)=U\Lambda^{1/2}\zeta$ and $Y'-E(Y')=U\Lambda^{1/2}\zeta'$. For any $v\in\mathcal{S}_X$,

$$\begin{aligned}\frac{1}{\sqrt{p}}v^T(Y-Y')&=\frac{1}{\sqrt{p}}v^TU\Lambda^{1/2}(\zeta-\zeta')\\ &=\sum_{i=1}^{m}\left(\frac{\lambda_i}{p}\right)^{1/2}v^Tu_i(\zeta_i-\zeta_i')+\frac{1}{\sqrt{p}}\sum_{i=m+1}^{p}\tau_iv^Tu_i(\zeta_i-\zeta_i'),\end{aligned} \tag{C.27}$$

where ζ_i and ζ_i' is ith component of ζ and ζ', respectively. With the aim of the second term in (C.27) converging to 0, we use a similar strategy as in Lemma

C.1. For any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\sum_{i=m+1}^{p}\tau_i v^T u_i(\zeta_i-\zeta_i')\right|>\epsilon\right)\le\frac{1}{p\epsilon^2}E\left(\left|\sum_{i=m+1}^{p}\tau_i v^T u_i(\zeta_i-\zeta_i')\right|^2\right)\\
=&\frac{1}{p\epsilon^2}E\left\{\sum_{i=m+1}^{p}\tau_i^2(v^Tu_i)^2(\zeta_i-\zeta_i')^2+\sum_{m+1\le i\neq\iota}\tau_i\tau_\iota(v^Tu_i)(v^Tu_\iota)(\zeta_i-\zeta_i')(\zeta_\iota-\zeta_\iota')\right\}\\
=&\frac{1}{p\epsilon^2}E\left\{\sum_{i=m+1}^{p}\tau_i^2(v^Tu_i)^2(\zeta_i-\zeta_i')^2\right\}\le\frac{M^2}{p\epsilon^2}E\left\{\sum_{i=m+1}^{p}(v^Tu_i)^2(\zeta_i-\zeta_i')^2\right\}\le\frac{2M^2}{p\epsilon^2}.
\end{aligned}\tag{C.28}$$

The first inequality is Chebyshev's inequality and the second equality holds because $\mathcal{X}$, Y and Y' are independent. The result in (C.28) gives that the second term in (C.27) converges to 0 in probability as well as in L^2. Therefore, we get

$$\frac{1}{\sqrt{p}}v^T(Y-Y')=\sum_{i=1}^{m}\sigma_i(\zeta_i-\zeta_i')v^Tu_i+o_P(1).$$

Since ζ and ζ' are invariant to p, above equation implies that $v^T(Y-Y')/\sqrt{p}\xrightarrow{P}0$ if and only if $v^Tu_i\xrightarrow{P}0$ for $i=1,\dots,m$, in other words, $\mathcal{A}=\mathcal{A}''$.

Now, we show that $\mathcal{A}=\mathcal{A}'$. First, $\mathcal{A}'\subset\mathcal{A}$ is clear due to Chebyshev's inequality. For the converse, it remains to show that $\mathrm{Var}\{p^{-1/2}v^T[Y-E\{Y|\pi(Y)=j\}]|\pi(Y)=j\}=E\{p^{-1}(v^T[Y-E\{Y|\pi(Y)=j\}])^2|\pi(Y)=j\}\to 0$ as $p\to\infty$ for $\{v\}\in\mathcal{A}$ and $j=1,2$. Since for both $j=1,2$,

$$p^{-1/2}v^T[Y-E\{Y|\pi(Y)=j\}]=\sum_{i=1}^{m}\left(\frac{\lambda_i}{p}\right)^{1/2}v^Tu_i\zeta_i+p^{-1/2}\sum_{i=m+1}^{p}\tau_i v^Tu_i\zeta_i,\tag{C.29}$$

if we follow the same logic in (C.28), the second term in (C.29) converges to 0 in L^2. Hence, it suffices to show that $E\{\lambda_i(v^Tu_i)^2\zeta_i^2/p\}\longrightarrow 0$ as $p\to\infty$ for $i=1,\dots,m$. Note that $\lambda_i/p=\sigma_i^2+O(p^{-1})$ and $(v^Tu_i)^2\le 1$. Therefore, Assumption 4 guarantees that for $i=1,\dots,m$, the sequence $\{\lambda_i(v^Tu_i)^2\zeta_i^2/p\}$, which converges to 0 in probability since $\{v\}\in\mathcal{A}=\mathcal{A}''$, has uniformly bounded second moments, so is uniformly integrable. Thus, with Vitali's convergence theorem, $p^{-1/2}\lambda_i^{1/2}v^Tu_i\zeta_i$ also converges to 0 in L^2 for $i=1,\dots,m$, which gives $\mathcal{A}\subset\mathcal{A}'$.

(ii) For an HDLSS-consistent $\hat{\alpha}$, write $\mathcal{B}_p=\mathrm{span}(v_{\hat{\alpha}})\oplus\mathrm{span}(\{\hat{u}_i\}_{i=m+1}^{n-2})$, an $(n-m-1)$-dimensional subspace of $S_X\subset\mathbb{R}^p$. For each p, let $\{v_{\hat{\alpha}},f_1,\dots,f_m\}$ forms an orthogonal basis of $\mathcal{S}$. Consequently, $\{v_{\hat{\alpha}},f_1,\dots,f_m,\hat{u}_{m+1},\dots,\hat{u}_{n-2}\}$ forms an orthogonal basis for $\mathcal{S}_X$. For a fixed $\{v\}\in\mathcal{A}$, write $v=a_0v_{\hat{\alpha}}+\sum_{i=1}^{m}a_if_i+\sum_{i=m+1}^{n-2}a_i\hat{u}_i$. Theorem 3.2 and Lemma C.3 (i) gives that ${v_{\hat{\alpha}}}^Tu_\iota=$

${v_{\hat\alpha}}^T P_{\mathcal{S}} u_\iota = {v_{\hat\alpha}}^T u_{\iota,\mathcal{S}} \xrightarrow{P} 0$ for $\iota = 1, \dots, m$ and $\hat u_i^T u_\iota \xrightarrow{P} 0$ for $i \geq m+1$ and $\iota = 1, \dots, m$, respectively. Since $v^T u_\iota \xrightarrow{P} 0$ for $\{v\} \in \mathcal{A}$ and $\iota = 1, \dots, m$, a_ι ($\iota = 1, \dots, m$) converges to 0 in probability. Consequently, $\|P_{\mathcal{B}_p} v\|^2 = \|v\|^2 - \sum_{\iota=1}^m a_\iota^2 \xrightarrow{P} 1$. Now, let $\{w\} \in \mathcal{B}_p$ with $w = P_{\mathcal{B}_p} v / \|P_{\mathcal{B}_p} v\| \in \mathcal{B}_p$ for all p. Then, $w^T v = \|P_{\mathcal{B}_p} v\| \xrightarrow{P} 1$, which gives that $\|w - v\| \xrightarrow{P} 0$ as $p \to \infty$. □

C.6. Proof of Theorem 3.4

Proof. Let $\beta = 1$ and $\hat\alpha$ be an HDLSS-consistent estimator of $-\tau^2/n$. We use same notation as in the proof of Lemma 3.3. For any $\{w\} \in \mathcal{A}$ such that $D(w)$ in (3.5) exists, the triangle inequality gives

$$|p^{-1/2} w^T \mu - D(w)| \leq p^{-1/2} |w^T \{Y_1 - E(Y_1)\}| + p^{-1/2} |w^T \{Y_2 - E(Y_2)\}| + o_P(1),$$

which implies that $|p^{-1/2} w^T \mu| \xrightarrow{P} D(w)$. Combining Lemma 3.3 and Lemma C.3 (i), for each $i = m+1, \dots, n-2$, $\{\hat u_i\}$ belongs to $\mathcal{A}$. Moreover, from (C.7), $D(\hat u_i) = 0$. Using the notation from the proof of Lemma 3.3, $w = a_0 v_{\hat\alpha} + \sum_{i=1}^m a_i f_i + \sum_{i=m+1}^{n-2} a_i \hat u_i$ while $a_i = o_P(1)$ for $i = 1, \dots, m$. From the triangle inequality,

$$\begin{aligned}
\left| \frac{1}{\sqrt{p}} v^T (Y_1 - Y_2) \right| &\leq \frac{1}{\sqrt{p}} \Big\{ \left| a_0 {v_{\hat\alpha}}^T (Y_1 - Y_2) \right| + \sum_{i=1}^m \left| a_i {f_i}^T (Y_1 - Y_2) \right| \\
&\quad + \sum_{i=m+1}^{n-2} \left| a_i \hat u_i^T (Y_1 - Y_2) \right| \Big\} \\
&= |a_0| D(v_{\hat\alpha}) + o_P(1) + \sum_{i=m+1}^{n-2} |a_i| D(\hat u_i) \\
&= |a_0| D(v_{\hat\alpha}) + o_P(1).
\end{aligned} \tag{C.30}$$

Since $|a_0| \leq 1$, the desired results are obtained immediately from (C.30). Here, the equality holds if and only if $a_0 \xrightarrow{P} 1$, which is also equivalent to that $\|w - v_{\hat\alpha}\| \xrightarrow{P} 0$.

For the second part of Theorem 3.4, recall $\tilde v_\alpha$ in (C.21) and let $v_{\hat\alpha}^T (Y - \bar X)/\sqrt{p} = M / \|\tilde v_{\hat\alpha}\|$ where $Y \in \mathcal{Y}$ and $M = \tilde v_{\hat\alpha}^T (Y - \bar X)/\sqrt{p}$. Since $1/\|\tilde v_{\hat\alpha}\| \xrightarrow{P} \gamma > 0$ in (C.26), it suffices to show that

$$M \xrightarrow{P} \begin{cases} \eta_2 (1 - k^2) \delta^2, & \text{for } \pi(Y) = 1; \\ -\eta_1 (1 - k^2) \delta^2, & \text{for } \pi(Y) = 2. \end{cases}$$

Assume that $\pi(Y) = 1$. The definition of $\tilde{v}_{\hat{\alpha}}$ in (C.21) gives

$$\begin{aligned} M = &\sum_{i=1}^{m} \frac{\hat{\alpha}}{\hat{\lambda}_i/p + \hat{\alpha}} \left(\frac{1}{\sqrt{p}} \hat{u}_i^T d \right) \left\{ \frac{1}{\sqrt{p}} \hat{u}_i^T (Y - \bar{X}) \right\} \\ &+ \frac{1}{\sqrt{p}} \| \hat{U}_2 \hat{U}_2^T d \| \left\{ \frac{1}{\sqrt{p}} w_{\mathrm{MDP}}^T (Y - \bar{X}) \right\}. \end{aligned} \tag{C.31}$$

We can derive the limit of M through simple arithmetics with the limits of the random variables in (C.31) which are already found in (C.13), (C.17), (C.20), Lemma C.2 and C.3 (ii). Specifically,

$$\begin{aligned} M \xrightarrow{P} &\sum_{i=1}^{m} \frac{-\tau^2}{\phi_i(\Omega)} \sum_{j=1}^{m} \sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega) + \tau^2}} \left\{ k_j \delta + \sigma_j \left(\bar{z}_{1,j} - \bar{z}_{2,j} \right) \right\} v_{ij}(\Omega) \\ &\times \sum_{j'=1}^{m} \sqrt{\frac{\phi_i(\Omega)}{\phi_i(\Omega) + \tau^2}} \left\{ \eta_2 k_{j'} \delta + \sigma_{j'} \left(\zeta_{j'} - \bar{z}_{j'} \right) \right\} v_{ij'}(\Omega) + \eta_2 (1 - k^2) \delta^2 \\ &+ \sum_{i=1}^{m} \frac{\tau^2}{\phi_i(\Omega) + \tau^2} \sum_{j=1}^{m} \left\{ k_j \delta + \sigma_j \left(\bar{z}_{1,j} - \bar{z}_{2,j} \right) \right\} v_{ij}(\Omega) \\ &\times \sum_{j'=1}^{m} \left\{ \eta_2 k_{j'} \delta + \sigma_{j'} \left(\zeta_{j'} - \bar{z}_{j'} \right) \right\} v_{ij'}(\Omega) \\ &= \eta_2 (1 - k^2) \delta^2 > 0, \end{aligned} \tag{C.32}$$

as $p \to \infty$. For $Y \in \mathcal{Y}$ with $\pi(Y) = 2$, $M \xrightarrow{P} -\eta_1(1 - k^2)\delta^2 < 0$ as $p \to \infty$ is similarly verified. □

C.7. Proof of Theorem 3.5

Proof. The correct classification rate of $\boldsymbol{\phi}_{\hat{\alpha}}$ is

$$\mathbb{P}\{\boldsymbol{\phi}_{\hat{\alpha}}(Y; \mathcal{X}) = \pi(Y)\} = \sum_{i=1}^{2} \mathbb{P}\{\boldsymbol{\phi}_{\hat{\alpha}}(Y; \mathcal{X}) = i | \pi(Y) = i\} \mathbb{P}\{\pi(Y) = i\}.$$

Here, the event $\{\boldsymbol{\phi}_{\hat{\alpha}}(Y; \mathcal{X}) = 1\}$ is equivalent to that $M \geq 0$ where $M = \tilde{v}_{\hat{\alpha}}^T (Y - \bar{X})/\sqrt{p}$. Since M converges to strictly positive quantity $\eta_2(1 - k^2)\delta^2$ for $Y \in \mathcal{Y}$ with $\pi(Y) = 1$, as shown in (C.32), we have $\mathbb{P}\{M \geq 0 | \pi(Y) = 1\} \to 1$. Similar argument can be used for $Y \in \mathcal{Y}$ with $\pi(Y) = 2$, leading that $\mathbb{P}\{M < 0 | \pi(Y) = 2\} \to 1$. Integrating these, we have $\mathbb{P}\{\boldsymbol{\phi}_{\hat{\alpha}}(Y; \mathcal{X}) = \pi(Y)\} \to 1$ as $p \to \infty$. □

C.8. Proof of Lemma 3.6

Proof. The performance of $\boldsymbol{\phi}_\alpha$ is analyzed through the inner product $N_\alpha = \tilde{v}_\alpha^T(Y - \bar{X})/\sqrt{p}$ where $Y \in \mathcal{Y}$. Inspecting (C.21), the quantity N_α becomes

$$\begin{aligned} N_\alpha = &\sum_{i=1}^m \frac{\alpha}{\hat{\lambda}_i/p + \alpha}\left(\frac{1}{\sqrt{p}}\hat{u}_i^T d\right)\left\{\frac{1}{\sqrt{p}}\hat{u}_i^T(Y - \bar{X})\right\} \\ &+ \frac{1}{\sqrt{p}}\|\hat{U}_2\hat{U}_2^T d\|\left\{\frac{1}{\sqrt{p}}w_{\mathrm{MDP}}^T(Y - \bar{X})\right\}. \end{aligned} \tag{C.33}$$

Now, we aim to obtain the limit of N_α.

(ii) Assume that $\beta = 1$ and $\pi(Y) = 1$. Utilizing the results in (C.13), (C.17), (C.20), Lemmas C.2 and C.3 (ii), we get the limit of N_α,

$$\begin{aligned} N_\alpha \xrightarrow{P} &\sum_{i=1}^m \frac{\alpha}{\tau^2/n + \phi_i(\Omega)/n + \alpha}\frac{\phi_i(\Omega)}{\tau^2 + \phi_i(\Omega)} \\ &\times \sum_{j=1}^m \{k_j\delta + \sigma_j(\bar{z}_{1,j} - \bar{z}_{2,j})\} v_{ij}(\Omega) \\ &\times \sum_{j'=1}^m \{\eta_2 k_{j'}\delta + \sigma_{j'}(\zeta_{j'} - \bar{z}_{j'})\} v_{ij'}(\Omega) + \eta_2(1 - k^2)\delta^2 \\ &+ \sum_{i=1}^m \frac{\tau^2}{\tau^2 + \phi_i(\Omega)} \sum_{j=1} \{k_j\delta + \sigma_j(\bar{z}_{1,j} - \bar{z}_{2,j})\} v_{ij}(\Omega) \\ &\times \sum_{j'=1}^m \{\eta_2 k_{j'}\delta + \sigma_{j'}(\zeta_{j'} - \bar{z}_{j'})\} v_{ij'}(\Omega) \\ &= \eta_2(1 - k^2)\delta^2 + \sum_{i=1}^m \frac{\tau^2 + n\alpha}{\phi_i(\Omega) + \tau^2 + n\alpha} \sum_{j=1} \{k_j\delta + \sigma_j(\bar{z}_{1,j} - \bar{z}_{2,j})\} v_{ij}(\Omega) \\ &\times \sum_{j'=1}^m \{\eta_2 k_{j'}\delta + \sigma_{j'}(\zeta_{j'} - \bar{z}_{j'})\} v_{ij'}(\Omega) \\ &\overset{d}{=} \eta_2(1 - k^2)\delta^2 + \left(n\alpha + \tau^2\right)(y - \bar{x})^T\left(\Omega + \left(n\alpha + \tau^2\right)I_m\right)^{-1}(\bar{x}_1 - \bar{x}_2) \\ &= \xi_\alpha + C_1. \end{aligned}$$

Here, $x_{i,j}$ and y are defined in Section 3.3 and $X \overset{d}{=} Y$ means that two random variables are equal in distribution. Since the convergence in probability implies the convergence in distribution, $\mathcal{P}_1(\alpha)$, the asymptotic correct classification rate of $\boldsymbol{\phi}_\alpha$ given that $\pi(Y) = 1$, becomes

$$\begin{aligned} \mathcal{P}_1(\alpha) &= \lim_{p\to\infty} \mathbb{P}\{\boldsymbol{\phi}_\alpha(Y; \mathcal{X}) = 1 | \pi(Y) = 1\} \\ &= \lim_{p\to\infty} \mathbb{P}\{N_\alpha \geq 0 | \pi(Y) = 1\} \end{aligned}$$

$$= \mathbb{P}\{\xi_\alpha + C_1 \geq 0 | \pi(Y) = 1\}.$$

Similar arguments applied to $\mathcal{P}_2(\alpha)$ leads that $N_\alpha \xrightarrow{P} \xi_\alpha - C_2$ and $\mathcal{P}_2(\alpha) = \mathbb{P}\{\xi_\alpha - C_2 < 0 | \pi(Y) = 2\}$.

(i) Most part before this proof covered the case when $\beta = 1$. Here, we assume $0 \leq \beta < 1$, which shows a sharp contrast with $\beta = 1$. We begin with re-calculating the limit of quantities composing N_α in (C.33). First, the leading eigenvalues no longer show any difference with the rest eigenvalues in the limit and $\hat{\lambda}_i / p \xrightarrow{P} \tau^2/n$ for $i = 1, \ldots, m$; see Lemma C.2. Hence, we exclude the case $\alpha = -\tau^2/n$ since otherwise the first term in (C.33) inflates. Also, Lemma C.3 (ii) gives that $\hat{u}_i^T d / \sqrt{p} \xrightarrow{P} 0$ for all i. We claim that

$$N_\alpha \xrightarrow{P} \begin{cases} \eta_2 \delta^2, & \pi(Y) = 1; \\ -\eta_1 \delta^2, & \pi(Y) = 2; \end{cases}, \tag{C.34}$$

as $p \to \infty$. If (C.34) holds, then we can conclude that $\mathcal{P}(\alpha) = 1$ for $\alpha \neq -\tau^2/n$ as in the proof of Theorem 3.5. In order to show (C.34), we need to obtain the limit of $p^{-1/2} \hat{u}_i^T (Y - \bar{X})$ $(i = 1, \ldots, m)$ and $p^{-1} (\hat{U}_2 \hat{U}_2^T d)^T (Y - \bar{X})$.

On $p^{-1/2} \hat{u}_i^T (Y - \bar{X})$. Assume that $\pi(Y) = 1$. The quantity $p^{-1/2} \hat{u}_i^T (Y - \bar{X})$ $(i = 1, \ldots, m)$ is analyzed through $\hat{U}_1^T (Y - \bar{X}) / \sqrt{p}$ by definition. In (C.16), note that the first and second equations hold regardless of β,

$$\frac{1}{\sqrt{p}} \hat{U}_1^T (Y - \bar{X}) = \frac{\eta_2}{\sqrt{p}} \hat{U}_1^T \mu + \frac{1}{\sqrt{p}} D^{-1} \hat{V}_1^T (I_n - J) Z^T \Lambda \left(\zeta - \frac{1}{n} Z 1_n \right).$$

Combining the results in (C.7) and Lemma C.1 for the case $0 \leq \beta < 1$, $\hat{U}_1^T (Y - \bar{X}) / \sqrt{p} \xrightarrow{P} 0$ as $p \to \infty$.

On $p^{-1} (\hat{U}_2 \hat{U}_2^T d)^T (Y - \bar{X})$. From the definition, $p^{-1} (\hat{U}_2 \hat{U}_2^T d)^T (Y - \bar{X}) = K_1 - K_2$ where $K_1 = d^T (Y - \bar{X}) / p$ and $K_2 = (\hat{U}_1 \hat{U}_1^T d)^T (Y - \bar{X}) / p$. The limit of K_1 can be obtained as in (C.18) utilizing Lemma C.1,

$$\begin{aligned} K_1 &= \frac{1}{p} \left(\mu + U \Lambda^{1/2} Z \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix} \right)^T \left\{ \eta_2 \mu + U \Lambda^{1/2} \left(\zeta - \frac{1}{n} Z 1_n \right) \right\} \\ &= \frac{1}{p} \left\{ \eta_2 \|\mu\|^2 + \eta_2 \mu^T U \Lambda^{1/2} Z \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix} + \mu^T U \Lambda^{1/2} \left(\zeta - \frac{1}{n} Z 1_n \right) \right. \\ &\quad \left. + \begin{bmatrix} \frac{1}{n_1} 1_{n_1} \\ -\frac{1}{n_2} 1_{n_2} \end{bmatrix}^T Z^T \Lambda \left(\zeta - \frac{1}{n} Z 1_n \right) \right\} \\ &\xrightarrow{P} \eta_2 \delta^2, \end{aligned}$$

as $p \to \infty$. As well, K_2 converges to 0 in probability since we have shown that the both quantities $\hat{U}_1^T d / \sqrt{p}$ and $\hat{U}_1^T (Y - \bar{X}) / \sqrt{p}$ converge to 0.

All in all, we have shown that $N_\alpha \xrightarrow{P} \eta_2\delta^2$ for $Y \in \mathcal{Y}$ with $\pi(Y) = 1$ and the other case when $\pi(Y) = 2$ can be done with very similar logic. □

C.9. Proof of Theorem 3.7

Proof. Assume $\beta = 1$. Lemma 3.6 tells that $\mathcal{P}_1(\alpha) = \mathbb{P}\{\xi_\alpha + C_1 \geq 0 | \pi(Y) = 1\}$ and $\mathcal{P}_2(\alpha) = \mathbb{P}\{\xi_\alpha - C_2 < 0 | \pi(Y) = 2\}$. It is immediate that if ξ_α given $\pi(Y)$ has its support on $(-\infty, \infty)$ for $\alpha \neq -\tau^2/n$, then $\mathcal{P}(\alpha)$ has its unique maximum 1 at $\alpha = -\tau^2/n$. Assume $\alpha \neq -\tau^2/n$ and let $\mathcal{I}$ be an any interval in $\mathbb{R}$ with a positive Lebesque measure. We claim that $\mathbb{P}(\xi_\alpha \in \mathcal{I}) > 0$. To see this, note that ξ_α given $x_{\mathrm{tr}} := \{x_{ij} : j = 1, \ldots, n_i, i = 1, 2\} \subset \mathbb{R}^n$ is a linear translation of y. Also, y equals to some linear translation of $z = (z_{(1)}, \ldots, z_{(m)})^T$ in distribution. Since $\{x : f_z(x) > 0\} = \mathbb{R}^m$, ξ_α given x_{tr} has its support on $(-\infty, \infty)$ with probability 1. Consequently, we have $\mathbb{P}(\xi_\alpha \in \mathcal{I} | x_{\mathrm{tr}}) > 0$, which implies that $\mathbb{P}(\xi_\alpha \in \mathcal{I}) = \mathbb{E}\{\mathbb{P}(\xi_\alpha \in \mathcal{I} | x_{\mathrm{tr}})\} > 0$. □

Appendix D: Additional numerical results

D.1. Additional figures

Figures D.1 to D.3 display the misclassification rates of binary classification of the MNIST and EMNIST datasets, for the cases where the number of leading components is set as $m = 1, 2, 3$. These figures are referenced in Section 4.2.

D.2. Model assumptions

For the MNIST data example, we check that Assumptions 1 and 2 are indeed satisfied. Checking the asymptotic assumptions is possible in this case, due to the following two reasons. First, dealing with varying dimensions is possible, since we can obtain random Fourier features of any dimension p. We have experimented with $p = 1000, 2000, 4000, 6000$ in Section 4.2. Second, since the data set consists of a sufficiently large number of sample (6,000 observations on average for each category), we may regard the sample estimate a close approximation of true parameter.

For each case of binary classification problems (total $\binom{10}{2} = 45$ cases), and for each dimension p, we have computed the size of the mean difference $\|\mu\| = \|\mu_1 - \mu_2\|$, and the eigenvalues λ_i of the within covariance matrix Σ. The mean difference μ and covariance matrix Σ are computed from the whole sample, and we treat as if there are the true parameters. (The experiment in Section 4.2 was conducted based on a random sample of size $n = 100$ or 200 from the whole sample.)

Assumption 2 is satisfied if $\|\mu\| = O(\sqrt{p})$ or, equivalently, $\|\mu\|^2 = O(p)$. In Fig. D.4, we confirm that $\|\mu\|^2$ is linear in p, and Assumption 2 is seemingly satisfied.

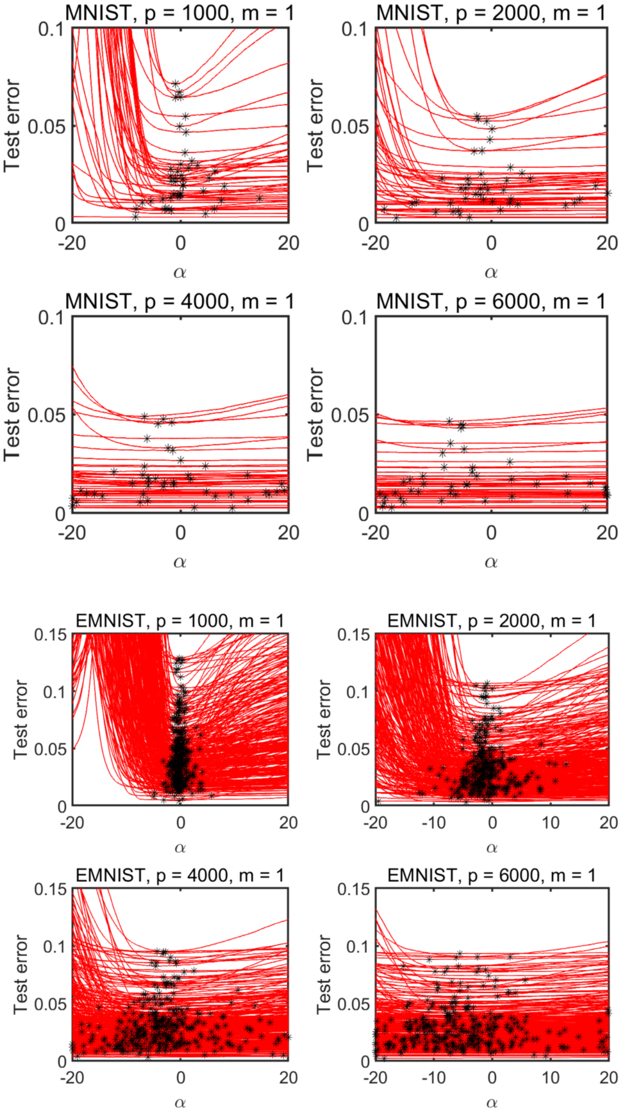


FIG D.1. *The $m = 1$ case. Misclassification rates of ϕ_α applied to two-group classifications of images of handwritten digits (MNIST) and alphabets (EMNIST).*

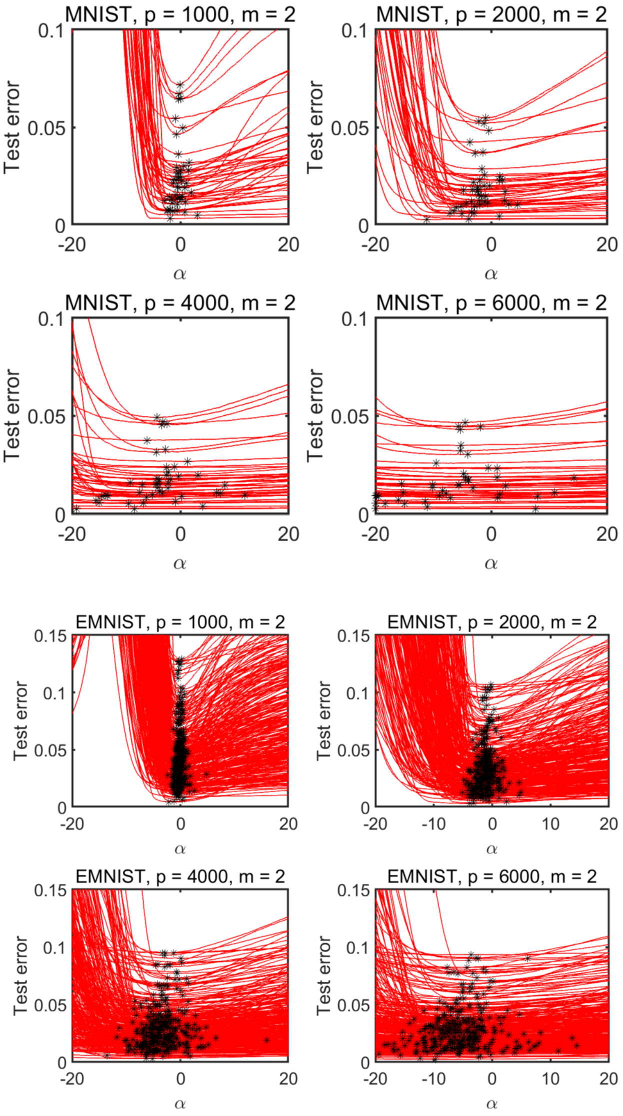


Fig D.2. *The $m = 2$ case. Misclassification rates of ϕ_α applied to two-group classifications of images of handwritten digits (MNIST) and alphabets (EMNIST).*

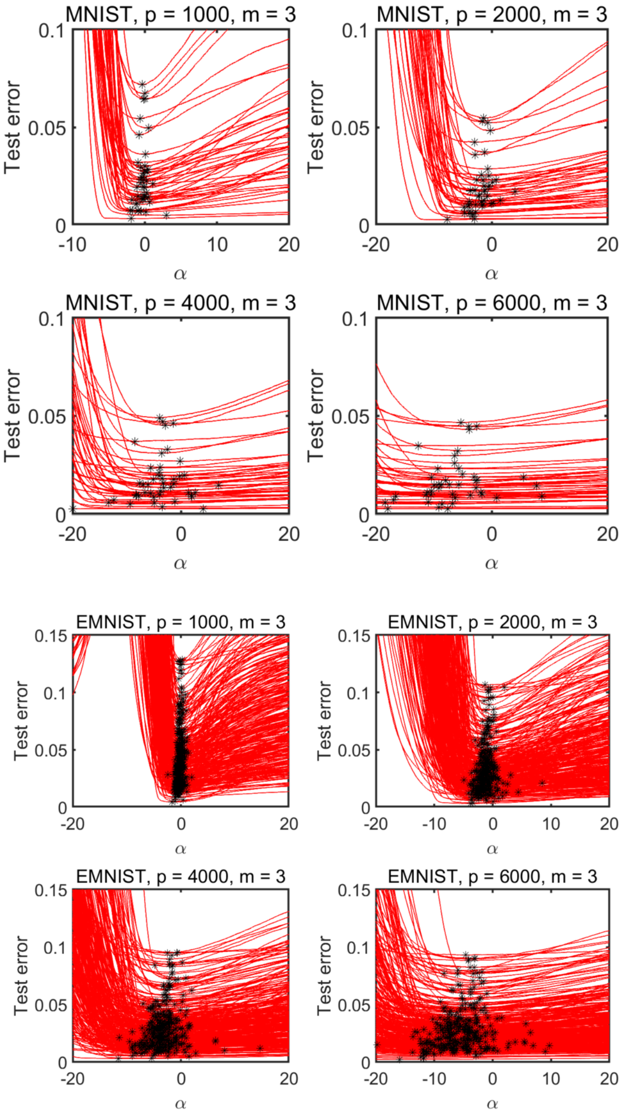


Fig D.3. *The $m = 3$ case. Misclassification rates of ϕ_α applied to two-group classifications of images of handwritten digits (MNIST) and alphabets (EMNIST).*

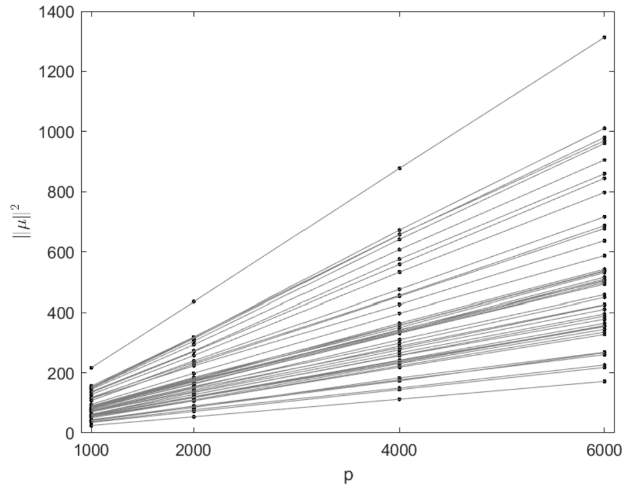


FIG D.4. $\|\mu\|^2$ *against the dimension* p *for all 45 binary problems of MNIST data. See text for the definition of* μ.

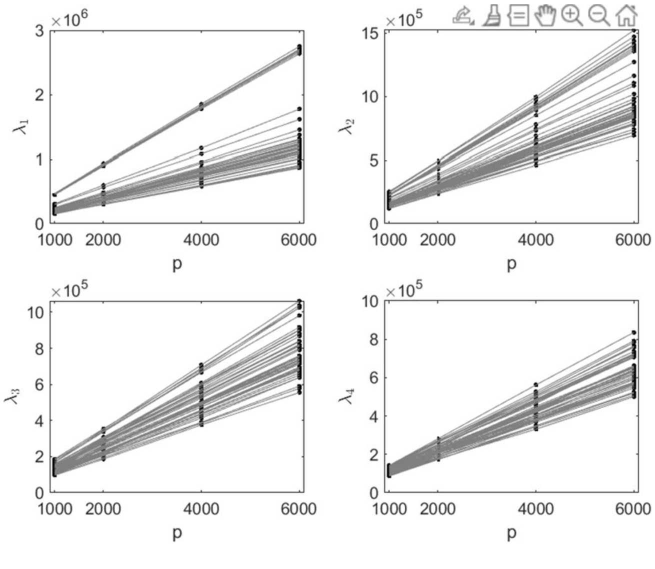


FIG D.5. *The eigenvalue against the dimension* p *for all 45 binary problems of MNIST data (plotted for the first four largest eigenvalues). See text for the definition of* λ_i.

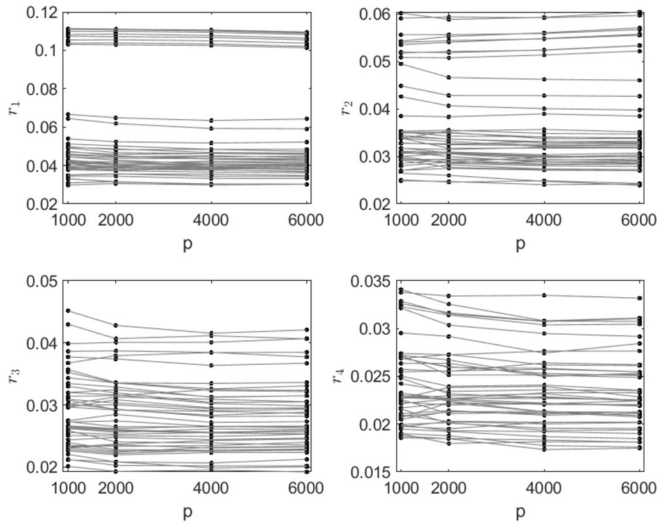


Fig D.6. *The eigenvalue ratio against the dimension p for all 45 binary problems of MNIST data (plotted for the first four largest eigenvalues). See text for the definition of r_i.*

Assumption 1 gives that the first m eigenvalues of Σ grow at the rate of p. Moreover, the ratio of λ_i over the total variance should be constant if the assumption is true. Define the ratio by

$$r_i := \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

If $\lambda_i = O(p)$ and r_i is constant over p, then the ith component may be considered as a leading component (or a "spike"). In Fig. D.5, we confirm that λ_i is indeed nearly linear in p for $i = 1, 2, 3, 4$; in Fig. D.6, we confirm that the ratio r_i is nearly constant for $i = 1, 2, 3, 4$.

To summarize, Figs. D.4-D.5 suggest that our key assumptions are satisfied for the MNIST dataset. Therefore, the numerical results on the classification in Section 4.2 should come at no surprise.

Acknowledgments

We would like to thank Editor, Associate Editor and anonymous reviewers whose comments and suggestions helped to improve and clarify our manuscript. We would also like to thank Mr. Taehyun Kim for constructive criticism of the manuscript.

References

Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika* **97** 254-259. MR2594434

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760-766. MR2410023

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* **96** 6745–6750.

Aoshima, M. and Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics* **71** 473–503. MR3968846

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics* **60** 4-19. MR3780619

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

Bartlett, P. L., Long, P. M., Lugosi, G. and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences* **117** 30063-30070. MR4263288

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

Bradley, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surveys* **2** 107-144. MR2178042

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R., Sugarbaker, D. J., Yeh, R.-F., Wiencke, J. K. and Kelsey, K. T. (2009). Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLOS Genetics* **5** e1000602.

Cohen, G., Afshar, S., Tapson, J. and van Schaik, A. (2017). EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 267-286. MR2899863

Di Pillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods* **5** 843-854. MR0518931

Fan, J., Wang, K., Zhong, Y. and Zhu, Z. (2021). Robust high dimensional factor models with applications to statistical machine learning. *Statistical Science* **36** 303-327. MR4255196

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

Friedman, J. H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association* **84** 165-175. MR0999675

Glaab, E., Bacardit, J., Garibaldi, J. M. and Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS one* **7** e39932.

Gravier, E., Pierron, G., Vincent-Salomon, A., gruel, N., Raynal, V., Savignoni, A., De Rycke, Y., Pierga, J.-Y., Lucchesi, C., Reyal, F., Fourquet, A., Roman-Roman, S., Radvanyi, F., Sastre-Garau, X., Asselain, B. and Delattre, O. (2010). A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer* **49** 1125–1125.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8** 86-100.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric Representation of High Dimension, Low Sample Size Data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 427–444. MR2155347

Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Hellton, K. H. and Thoresen, M. (2017). When and Why are Principal Component Scores a Good Tool for Visualizing High-dimensional Data? *Scandinavian Journal of Statistics* **44** 581–597. MR3687964

Holzmüller, D. (2020). On the Universality of the Double Descent Peak in Ridgeless Regression. *arXiv preprint arXiv:2010.01851*.

Ishii, A., Yata, K. and Aoshima, M. (2019). Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model. *Japanese Journal of Statistics and Data Science* **2** 105-128. MR3969141

Jeffery, I. B., Higgins, D. G. and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics* **7** 1–16.

Jung, S. (2018). Continuum directions for supervised dimension reduction. *Computational Statistics & Data Analysis* **125** 27 - 43. MR3800144

Jung, S., Lee, M. H. and Ahn, J. (2018). On the number of principal components in high dimensions. *Biometrika* **105** 389-402. MR3804409

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics* **37** 4104-4130. MR2572454

Kobak, D., Lomond, J. and 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** 1-16. MR4209455

Kolmogorov, A. N. and Rozanov, Y. A. (1960). On Strong Mixing Conditions for Stationary Gaussian Processes. *Theory of Probability & Its Applications* **5** 204-208. MR0133175

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

Lee, M. H., Ahn, J. and Jeon, Y. (2013). HDLSS Discrimination With Adaptive Data Piling. *Journal of Computational and Graphical Statistics* **22** 433-

451. MR3173723

Naderi, A., Teschendorff, A., Barbosa-Morais, N., Pinder, S., Green, A., Powe, D., Robertson, J., Aparicio, S., Ellis, I., Brenton, J. et al. (2007). A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene* **26** 1507–1516.

Passemier, D. and Yao, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *J. Multivar. Anal.* **127** 173–183. MR3188885

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J. and Marron, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association* **105** 401–414. MR2656058

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems* **20** 1177-1184.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S. et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine* **8** 68–74.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell* **1** 203–209.

Tsigler, A. and Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

Wu, D. and Xu, J. (2020). On the Optimal Weighted l_2 Regularization in Overparameterized Linear Regression. *arXiv preprint arXiv:2006.05800*.

Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* **105** 193 - 215. MR2877512

Yata, K. and Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics* **47** 899-921. MR4157163