The Annals of Statistics
2023, Vol. 51, No. 3, 1232–1257
https://doi.org/10.1214/23-AOS2289

OPTIMAL DISCRIMINANT ANALYSIS IN HIGH-DIMENSIONAL LATENT FACTOR MODELS

BY XIN BING[1,a] AND MARTEN WEGKAMP[2,b]

[1]*Department of Statistical Sciences, University of Toronto,* [a]*xin.bing@utoronto.ca*
[2]*Department of Mathematics & Department of Statistics and Data Science, Cornell University,* [b]*mhw73@cornell.edu*

In high-dimensional classification problems, a commonly used approach is to first project the high-dimensional features into a lower-dimensional space, and base the classification on the resulting lower-dimensional projections. In this paper, we formulate a latent-variable model with a hidden low-dimensional structure to justify this two-step procedure and to guide which projection to choose. We propose a computationally efficient classifier that takes certain principal components (PCs) of the observed features as projections, with the number of retained PCs selected in a data-driven way. A general theory is established for analyzing such two-step classifiers based on any projections. We derive explicit rates of convergence of the excess risk of the proposed PC-based classifier. The obtained rates are further shown to be optimal up to logarithmic factors in the minimax sense. Our theory allows the lower dimension to grow with the sample size and is also valid even when the feature dimension (greatly) exceeds the sample size. Extensive simulations corroborate our theoretical findings. The proposed method also performs favorably relative to other existing discriminant methods on three real data examples.

1. Introduction. In high-dimensional classification problems, a widely used technique is to first project the high-dimensional features into a lower-dimensional space, and base the classification on the resulting lower-dimensional projections [3, 11, 17, 22, 24, 30, 31, 36–38, 41, 42]. Despite having been widely used for years, theoretical understanding of this approach is scarce, and what kind of low-dimensional projection to choose remains unknown. In this paper, we formulate a latent-variable model with a hidden low-dimensional structure to justify the two-step procedure that takes leading principal components of the observed features as projections.

Concretely, suppose our data consists of independent copies of the pair (X, Y) with features $X \in \mathbb{R}^p$ according to

$$X = AZ + W \tag{1.1}$$

and labels $Y \in \{0, 1\}$. Here, A is a deterministic, unknown $p \times K$ loading matrix, $Z \in \mathbb{R}^K$ are unobserved, latent factors and W is random noise. We assume that:

(i) W is independent of both Z and Y,
(ii) $\mathbb{E}[W] = \mathbf{0}_p$,
(iii) A has rank K.

This mathematical framework allows for a substantial dimension reduction in classification for $K \ll p$. Indeed, in terms of the Bayes' misclassification errors, we prove in Lemma 1 of

Received August 2022; revised March 2023.
MSC2020 subject classifications. 62H12, 62J07.
Key words and phrases. High-dimensional classification, latent factor model, principal component regression, dimension reduction, discriminant analysis, optimal rate of convergence.

Section 2.1 the inequality

$$R_x^* := \inf_g \mathbb{P}\{g(X) \neq Y\} \geq R_z^* := \inf_h \mathbb{P}\{h(Z) \neq Y\}, \tag{1.2}$$

that is, it is easier to classify in the latent space $\mathbb{R}^K$ than in the observed feature space $\mathbb{R}^p$. In this work, we further assume that:

(iv) Z is a mixture of two Gaussians

$$Z|Y = k \sim N_K(\alpha_k, \quad \Sigma_{Z|Y}), \mathbb{P}(Y = k) = \pi_k, \quad k \in \{0, 1\} \tag{1.3}$$

with different means $\alpha_0 := \mathbb{E}[Z|Y = 0]$ and $\alpha_1 := \mathbb{E}[Z|Y = 1]$, but with the same covariance matrix

$$\Sigma_{Z|Y} := \mathrm{Cov}\,(Z|Y = 0) = \mathrm{Cov}\,(Z|Y = 1), \tag{1.4}$$

assumed to be strictly positive definite.

We emphasize that the distributions of X given Y are not necessarily Gaussian as the distribution of W could be arbitrary.

Within the above modeling framework, parameters related with the moments of X and Y, such as π_k, $\mathbb{E}[X|Y]$ and $\mathrm{Cov}(X|Y)$, are identifiable, while A, $\Sigma_{Z|Y}$, α_k and $\Sigma_W := \mathrm{Cov}(W)$ are not. For instance, we can always replace Z by $Z' = QZ$ for any invertible $K \times K$ matrix Q and write $\alpha_k' = Q\alpha_k$, $\Sigma_{Z|Y}' = Q\Sigma_{Z|Y}Q^\top$ and $A' = AQ^{-1}$. Since we focus on classification, there is no need to impose any conditions on the latter group of parameters that render them identifiable. Although our discussion throughout this paper is based on a fixed notation of A, $\Sigma_{Z|Y}$, Σ_W and α_k, it should be understood that our results are valid for all possible choices of these parameters such that model (1.1) and (1.3) holds, including submodels under which such parameters are (partially) identifiable.

Our goal is to construct a classification rule $\widehat{g}_x : \mathbb{R}^p \to \{0, 1\}$ based on the training data $\boldsymbol{D} := \{\boldsymbol{X}, \boldsymbol{Y}\}$ that consists of independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ from model (1.1) and (1.3) such that the resulting rule has small missclassification error $\mathbb{P}\{\widehat{g}_x(X) \neq Y\}$ for a new pair of (X, Y) from the same model that is independent of $\boldsymbol{D}$. In this paper, we are particularly interested in $\widehat{g}_x$ that is linear in X, motivated by the fact that the restriction of equal covariance in (1.4) leads to a Bayes rule that is linear in Z when we observe Z (see display (1.6) below).

Linear classifiers have been popular for decades, especially in high-dimensional classification problems, due to their interpretability and computational simplicity. One strand of the existing literature imposes sparsity on the coefficients $\beta \in \mathbb{R}^p$ in linear classifiers $g(x) = \mathbb{1}\{\beta^\top x + \beta_0 \geq 0\}$ for large p ($p \geq n$); see, for instance, [18, 19, 27, 40, 43, 48, 52] for sparse linear discriminant analysis (LDA) and [47, 51] for sparse support vector machines. For instance, in the classical LDA-setting, when X itself is a mixture of Gaussians

$$X|Y = k \sim N_p(\mu_k, \Sigma), \qquad \mathbb{P}(Y = k) = \pi_k, \quad k \in \{0, 1\} \tag{1.5}$$

with Σ strictly positive definite, the Bayes classifier is linear with p-dimensional vector $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$. Sparsity of β is then a reasonable assumption when Σ is close to diagonal, so that sparsity of β gets translated to that of the difference between the mean vectors $\mu_1 - \mu_0$. However, in the high-dimensional regime, many features are highly correlated and any sparsity assumption on β is no longer intuitive and becomes in fact questionable. This serves as a main motivation for this work, in which we study a class of linear classifiers that no longer requires the sparsity assumption on β, for neither construction of the classifier, nor its analysis.

1.1. *Contributions*. We summarize our contributions below.

1.1.1. *Minimax lower bounds of rate of convergence of the excess risk*. Our first contribution in this paper is to establish minimax lower bounds of rate of convergence of the excess risk for any classifier under model (1.1) and (1.3). The excess risk is defined relative to R_z^* in (1.2), which we view as a more natural benchmark than R_x^* because our proposed classifier is designed to adapt to the underlying low-dimensional structure in (1.1). The relation in (1.2) suggests R_z^* is also a more ambitious benchmark than R_x^*.

Since the gap between R_x^* and R_z^* quantifies the irreducible error for not observing Z, we start in Lemma 2 of Section 2.1 by characterizing how $R_x^* - R_z^*$ depends on $\xi^* = \lambda_K(A\Sigma_{Z|Y}A^\top)/\lambda_1(\Sigma_W)$, the signal-to-noise ratio for predicting Z from X (conditioned on Y), and $\Delta^2 = (\alpha_1 - \alpha_0)^\top \Sigma_{Z|Y}^{-1}(\alpha_1 - \alpha_0)$, the Mahalanobis distance between random vectors $Z|Y = 1$ and $Z|Y = 0$. Interestingly, it turns out that $R_x^* - R_z^*$ is small when either ξ^* or Δ is large, a phenomenon that is different from the setting when Y is linear in Z. Indeed, for the latter case, the excess risk of predicting Y by using the best linear predictor of X relative to the risk of predicting Y from $\mathbb{E}[Y|Z]$ is small only when ξ^* is large [13].

In Theorem 3 of Section 2.2, we derive the minimax lower bounds of the excess risk for any classifier with explicit dependency on the signal-to-noise ratio ξ^*, the separation distance Δ, the dimensions K and p and the sample size n. Our results also fully capture the phase transition of the excess risk as the magnitude of Δ varies. Specifically, when Δ is of constant order, the established lower bounds are

$$(\omega_n^*)^2 = \frac{K}{n} + \frac{\Delta^2}{\xi^*} + \frac{\Delta^2}{\xi^*}\frac{p}{\xi^* n}.$$

The first term is the optimal rate of the excess risk even when Z were observable; the second term corresponds to the irreducible error of not observing Z in $R_x^* - R_z^*$ and the last term reflects the minimal price to pay for estimating the column space of A. When $\Delta \to \infty$ as $n \to \infty$, the lower bounds become $(\omega_n^*)^2 \exp(-\Delta^2/8)$ and get exponentially faster in Δ^2. When $\Delta \to 0$ as $n \to \infty$, the lower bounds get slower as $\omega_n^* \min\{\omega_n^*/\Delta, 1\}$, implying a more difficult scenario for classification. In Section 5.3, the lower bounds are further shown to be tight in the sense that the excess risk of the proposed PC-based classifiers have a matching upper bound, up to some logarithmic factors.

To the best of our knowledge, our minimax lower bounds are both new in the literature of factor models and the classical LDA. In the factor model literature, even in linear factor regression models, there is no known minimax lower bound of the prediction risk with respect to the quadratic loss function. In the LDA literature, our results cover the minimax lower bound of the excess risk in the classical LDA as a special case and are the first to fully characterize the phase transition in Δ (see Remark 5 for details). The analysis of establishing Theorem 3 is highly nontrivial and encounters several challenges. Specifically, since the excess risk is not a semidistance, as required by the standard techniques of proving minimax lower bounds, the first challenge is to develop a reduction scheme based on a surrogate loss function that satisfies a local triangle inequality-type bound. The second challenge of our analysis is to allow a fully nondiagonal structure of $\mathrm{Cov}(X|Y)$ under model (1.1), as opposed to the existing literature on the classical LDA that assumes $\mathrm{Cov}(X|Y)$ to be diagonal or even proportional to the identity matrix. To characterize the effect of estimating the column space of A on the excess risk in deriving the third term of the lower bounds, our proof is based on constructing a suitable subset of the parameter space via the hypercube construction that is used for proving the optimal rates of the sparse PCA [50] (see the paragraph after Theorem 3 for a full discussion). Since the statistical distance (such as the KL-divergence) between thus constructed hypotheses could diverge as $p/n \to \infty$, this leads to the third challenge of providing a meaningful and sharp lower bound that is valid for both $p < n$ and $p > n$.

1.1.2. *A general two-step classification approach and the PC-based classifier.* Our second contribution in this paper is to propose a computationally efficient linear classifier in Section 3.2 that uses leading principal components (PCs) of the high-dimensional feature, with the number of retained PCs selected in a data-driven way. This PC-based classifier is one instance of a general two-step classification approach proposed in Section 3.1. To be clear, it differs from naively applying standard LDA, using plug-in estimates of the Bayes rule, on the leading PCs.

To motivate our approach, suppose that the factors Z were observable. Then the optimal Bayes rule is to classify a new point $z \in \mathbb{R}^K$ as

$$g_z^*(z) = \mathbb{1}\{z^\top \eta + \eta_0 \geq 0\}, \tag{1.6}$$

where

$$\eta = \Sigma_{Z|Y}^{-1}(\alpha_1 - \alpha_0), \qquad \eta_0 = -\frac{1}{2}(\alpha_0 + \alpha_1)^\top \eta + \log \frac{\pi_1}{\pi_0}. \tag{1.7}$$

This rule is optimal in the sense that it has the smallest possible misclassification error. Our approach in Section 3.1 utilizes an intimate connection between the linear discriminant analysis and regression to reformulate the Bayes rule $g_z^*(z)$ as $\mathbb{1}\{z^\top \beta + \beta_0 \geq 0\}$ with $\beta = \Sigma_Z^{-1} \operatorname{Cov}(Z, Y)$ (and β_0 is given in (3.1) of Section 3). The key difference is the use of the *unconditional* covariance matrix Σ_Z, as opposed to the *conditional* one $\Sigma_{Z|Y}$ in (1.7). As a result, β can be interpreted as the coefficient of regressing Y on Z, suggesting to estimate $z^\top \beta$ by $z^\top (\boldsymbol{Z}^\top \Pi_n \boldsymbol{Z})^+ \boldsymbol{Z}^\top \Pi_n \boldsymbol{Y}$ via the method of least squares, again, in case $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^{n \times K}$ and $z \in \mathbb{R}^K$ had been observed. Here, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \{0, 1\}^n$, $\Pi_n = \boldsymbol{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^\top$ is the centering projection matrix and M^+ denotes the Moore–Penrose inverse of any matrix M throughout of this paper.

Since we only have access to $x \in \mathbb{R}^p$, a realization of X, $\boldsymbol{X} = [X_1 \cdots X_n]^\top \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y} \in \{0, 1\}^n$, it is natural to estimate the span of z by $B^\top x$ and to predict the span of $\Pi_n \boldsymbol{Z}$ by $\Pi_n \boldsymbol{X} B$, for some appropriate matrix B. This motivates us to estimate the inner-product $z^\top \beta$ by

$$\big(B^\top x\big)^\top \big(B^\top \boldsymbol{X}^\top \Pi_n \boldsymbol{X} B\big)^+ B^\top \boldsymbol{X}^\top \Pi_n \boldsymbol{Y} := x^\top \widehat{\theta}. \tag{1.8}$$

By using a plug-in estimator $\widehat{\beta}_0$ of β_0, the resulting rule $\widehat{g}_x(x) = \mathbb{1}\{x^\top \widehat{\theta} + \widehat{\beta}_0 \geq 0\}$ is a general two-step, regression-based classifier and the choice of B is up to the practitioner.

In this paper, we advocate the choice $B = \boldsymbol{U}_r \in \mathbb{R}^{p \times r}$ where $\boldsymbol{U}_r$ contains the first r right-singular vectors of $\Pi_n \boldsymbol{X}$, such that the projections $\Pi_n \boldsymbol{X} B$ become the first r principal components of $\boldsymbol{X}$. Intuitively, this method has promise as [45] proves that when r is chosen as K, the projection $\Pi_n \boldsymbol{X} \boldsymbol{U}_K$ accurately predicts the span of $\Pi_n \boldsymbol{Z}$ under model (1.1). Since in practice K is oftentimes unknown, we further use a data-driven selection of K in Section 3.3 to construct our final PC-based classifier. The proposed procedure is computationally efficient. Its only computational burden is that of computing the singular value decomposition (SVD) of $\boldsymbol{X}$. Guided by our theory, we also discuss a cross-fitting strategy in Section 3.2 that improves the PC-based classifier by removing the dependence from using the data twice (one for constructing $\boldsymbol{U}_r$ and one for computing $\widehat{\theta}$ in (1.8)) when $p > n$ and the signal-to-noise ratio ξ^* is weak.

Retaining only a few principal components of the observed features and using them in subsequent regressions is known as principal component regression (PCR) [45]. It is a popular method for predicting $Y \in \mathbb{R}$ from a high-dimensional feature vector $X \in \mathbb{R}^p$ when both X and Y are generated via a low-dimensional latent factor Z. Most of the existing literature analyzes the performance of PCR when both Y and X are linear in Z, for instance, [6, 7, 13, 32, 45, 46], just to name a few. When Y is not linear in Z, little is known. An exception is

[29], which studies the model $Y = h(\xi_1 Z, \ldots, \xi_q Z; \varepsilon)$ and $X = AZ + W$ for some unknown general link function $h(\cdot)$. Their focus is only on estimation of $\xi_1, \ldots, \xi_q$, the sufficient predictive indices of Y, rather than analysis of the risk of predicting Y. As $\mathbb{E}[Y|Z]$ is not linear in Z under our models (1.1) and (1.3), to the best of our knowledge, analysis of the misclassifcation error under models (1.1) and (1.3) for a general linear classifier has not been studied elsewhere.

1.1.3. *A general strategy of analyzing the excess risk of $\widehat{g}_x$ based on any matrix B.* Our third contribution in this paper is to provide a general theory for analyzing the excess risk of the type of classifiers $\widehat{g}_x$ that uses a generic matrix B in (1.8). In Section 4, we state our result in Theorem 5, a general bound for the excess risk of the classifier $\widehat{g}_x$ based on a generic matrix B. It depends on (i) how well we estimate $z^\top \beta + \beta_0$ and (ii) a margin condition on the conditional distributions $Z|Y = k$, $k \in \{0, 1\}$, nearby the hyperplane $\{z|z^\top \beta + \beta_0 = 0\}$. This is a different approach than the usual one in the literature [26] that provides bounds on the excess risk $\mathbb{P}\{\widehat{g}(X) \neq Y|\boldsymbol{D}\} - R_z^*$ of a classifier $\widehat{g} : \mathbb{R}^p \to \{0, 1\}$ by the expression $2\mathbb{E}[|\eta(Z) - 1/2|\mathbb{1}\{\widehat{g}(X) \neq g_z^*(Z)\}|\boldsymbol{D}]$, with $\eta(z) = \mathbb{P}(Y = 1|Z = z)$, and involves analyzing the behavior of $\eta(Z)$ near $1/2$ (see our detailed discussion in Remark 7). The analysis of Theorem 5 is powerful in that it can easily be generalized to any distribution of $Z|Y$, as explained in Remark 8. Our second main result in Theorem 7 of Section 4 provides explicit rates of convergence of the excess risk of $\widehat{g}_x$ for a generic B and clearly delineates three key quantities that need to be controlled as introduced therein. The established rates of convergence reveal the same phase transition in Δ from the lower bounds. It is worth mentioning that the analysis of Theorem 7 is more challenging under models (1.1) and (1.3) than the classical LDA setting (1.5) in which the excess risk of any linear classifier in X has a closed-form expression.

1.1.4. *Optimal rates of convergence of the PC-based classifier.* Our fourth contribution is to apply the general theory in Section 4 to analyze the PC-based classifiers. Consistency of our proposed estimator of K is established in Theorem 8 of Section 5.1. In Theorem 9 of Section 5.2, we derive explicit rates of convergence of the excess risk of the PC-based classifier that uses $B = \boldsymbol{U}_K$. The obtained rate of convergence exhibits an interesting interplay between the sample size n and the dimensions K and p through the quantities K/n, ξ^* and Δ. Our analysis also covers the low signal setting $\Delta = o(1)$, a regime that has not been analyzed even in the existing literature of classical LDA. Our theoretical results are valid for both fixed and growing K and are also valid even when p is much lager than n. In Theorem 10 of Section 5.2, we also show that a PC-based LDA that uses either auxiliary data or sample splitting could surprisingly yield faster rates of convergence of the excess risk by removing the dependence between $\boldsymbol{U}_K$ and $\boldsymbol{X}$. These faster rates are further shown to be minimax optimal, up to a logarithmic factor, in Corollary 11 of Section 5.3. The benefit of using auxiliary data or sample splitting has also been recognized in other problems, such as the problem of estimating the optimal instrument in sparse high-dimensional instrumental variable model [10] and the problem of inference on a low-dimensional parameter in the presence of high-dimensional nuisance parameters [21].

1.1.5. *Extension to multiclass classification.* Our fifth contribution is to extend the general two-step classification procedure in Section 3 to handle multiclass classification problems in Section 8. Rates of convergence of the excess risk of the proposed multiclass classifier are derived in Theorem 12. PC-based classifiers are analyzed subsequently in Corollary 13. Our theory is the first to explicitly characterize dependence of the excess risk on the number of classes, and to cover the weak separation case when $\Delta \to 0$.

The paper is organized as follows. In Section 2.1, we provide an oracle benchmark that quantifies the excess risk of the optimal classifier based on X. We state the minimax lower

bounds of the excess risk for any classifier in Section 2.2. In Section 3, we present a connection between the linear discriminant classifier by using Z and regression of Y onto Z. This key observation leads to our proposed PC-based classifier. Furthermore, we propose a data-driven selection of the number of retained principal components. A general theory is stated in Section 4 for analyzing the excess risk of the classifier $\widehat{g}_x$ that uses any B for the estimate $\widehat{\theta}$ in (1.8). In Section 5, we apply the general result to analyze the PC-based classifiers. Main simulation results are presented in Section 6 and a real data analysis is given in Section 7. Extension to multiclass classification is studied in Section 8. All the proofs and additional simulation results are deferred to the Appendix [15].

Notation: We use the common notation $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ for the standard normal density, and denote by $\Phi(x) = \int \varphi(t)\mathbb{1}\{t \le x\}\,\mathrm{d}t$ its c.d.f. For any positive integer d, we write $[d] := \{1, \ldots, d\}$. For any vector v, we use $\|v\|_q$ to denote its ℓ_q norm for $0 \le q \le \infty$. We also write $\|v\|_Q^2 = v^\top Q^{-1} v$ for any commensurate, invertible square matrix Q. For any real-valued matrix $M \in \mathbb{R}^{r\times q}$, we use M^+ to denote the Moore–Penrose inverse of M, and $\sigma_1(M) \ge \sigma_2(M) \ge \cdots \ge \sigma_{\min(r,q)}(M)$ to denote the singular values of M in nonincreasing order. We define the operator norm $\|M\|_{\mathrm{op}} = \sigma_1(M)$. For a symmetric positive semidefinite matrix $Q \in \mathbb{R}^{p\times p}$, we use $\lambda_1(Q) \ge \lambda_2(Q) \ge \cdots \ge \lambda_p(Q)$ to denote the eigenvalues of Q in nonincreasing order. We write $Q \succ 0$ if Q is strictly positive definite. For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists some constant C such that $a_n \le C b_n$. The notation $a_n \asymp b_n$ stands for $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two numbers a and b, we write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. We use $\boldsymbol{I}_d$ to denote the $d \times d$ identity matrix and use $\mathbf{1}_d$ ($\mathbf{0}_d$) to denote the vector with all ones (zeroes). For $d_1 \ge d_2$, we use $\mathcal{O}_{d_1\times d_2}$ to denote the set of all $d_1 \times d_2$ matrices with orthonormal columns. Lastly, we use c, c', C, C' to denote positive and finite absolute constants that unless otherwise indicated can change from line to line.

2. Excess risk and its minimax optimal rates of convergence.

We start in Section 2.1 by introducing the oracle benchmark relative to which the excess risk is defined. Minimax optimal rates of convergence of the excess risk are derived in Section 2.2.

2.1. *Oracle benchmark.*

Since our goal is to predict the Bayes rule $\mathbb{1}\{z^\top \eta + \eta_0 \ge 0\}$ under model (1.3), it is natural to choose the oracle risk R_z^* in (1.2) as our benchmark, as opposed to R_x^*. Furthermore, we always have the explicit expression

$$R_z^* = 1 - \pi_1 \Phi\left(\frac{\Delta}{2} + \frac{\log\frac{\pi_1}{\pi_0}}{\Delta}\right) - \pi_0 \Phi\left(\frac{\Delta}{2} - \frac{\log\frac{\pi_1}{\pi_0}}{\Delta}\right); \tag{2.1}$$

see, for instance, [35], Section 8.3, pages 241–244. Here,

$$\Delta^2 := (\alpha_0 - \alpha_1)^\top \Sigma_{Z|Y}^{-1} (\alpha_0 - \alpha_1) \tag{2.2}$$

is the Mahalanobis distance between the conditional distributions $Z|Y = 1 \sim N_K(\alpha_1, \Sigma_{Z|Y})$ and $Z|Y = 0 \sim N_K(\alpha_0, \Sigma_{Z|Y})$. In particular, when $\pi_0 = \pi_1$, the expression in (2.1) simplifies to $R_z^* = 1 - \Phi(\Delta/2)$.

REMARK 1. It is immediate from (2.1) that $\Delta \to \infty$ implies $R_z^* \to 0$. The case of zero Bayes error R_z^* represents the easiest classification problem and we can expect fast rates of the excess risk. If $\Delta \to 0$, the Bayes risk R_z^* converges to $\min\{\pi_0, \pi_1\}$. When $\pi_0 = \pi_1 = 1/2$, the limit reduces to random guessing, which represents the hardest classification problem and slow rates are to be expected. When $\pi_0 \ne \pi_1$, we can expect fast rates, too, since the asymptotic Bayes rule always votes for the same label, to wit, the one with the largest unconditional probability. Thus, in a way, $\Delta \asymp 1$ is the most interesting case to investigate.

The lemma below shows that $R_x^* \geq R_z^*$, implying that R_z^* is also an ambitious benchmark.

LEMMA 1. *Under model* (1.1) *and* (i)–(iii), *we have*

$$R_x^* = \inf_{g:\mathbb{R}^p \to \{0,1\}} \mathbb{P}\big\{g(AZ+W) \neq Y\big\} \geq R_z^* = \inf_{h:\mathbb{R}^K \to \{0,1\}} \mathbb{P}\big\{h(Z) \neq Y\big\}.$$

PROOF. See Appendix A.1.1. □

If $W = \mathbf{0}_p$, the inequality in Lemma 1 obviously becomes an equality. More generally, if the signal for predicting Z from X under model (1.1) is large, we expect the gap between R_x^* and R_z^* to be small. To characterize such dependence, we introduce the following parameter space of $\theta := (A, \Sigma_{Z|Y}, \Sigma_W, \alpha_1, \alpha_0, \pi_1, \pi_0)$:

$$\Theta(\lambda, \sigma, \Delta) = \big\{\theta : \lambda_j(\Sigma_W) \asymp \sigma^2, \forall j \in [p], \lambda_k\big(A\Sigma_{Z|Y}A^\top\big) \asymp \lambda, \forall k \in [K], \pi_0 = \pi_1\big\} \tag{2.3}$$

and recall Δ from (2.2). For any $\theta \in \Theta(\lambda, \sigma, \Delta)$, the quantity λ/σ^2 can be treated as the signal-to-noise ratio for predicting Z from X given Y under model (1.1). The following lemma shows how the gap between R_x^* and R_z^* depends on λ/σ^2 and Δ in the special case $W \sim N_p(\mathbf{0}_p, \Sigma_W)$.

LEMMA 2. *Under model* (1.1) *and* (i)–(iv), *suppose* $W \sim N_p(\mathbf{0}_p, \Sigma_W)$ *with* $\Sigma_W \succ 0$. *For any* $\theta \in \Theta(\lambda, \sigma, \Delta)$, *we have*

$$\frac{\Delta}{1+(\lambda/\sigma^2)} \exp\left\{-\frac{\Delta^2}{8}\right\} \lesssim R_x^* - R_z^* \lesssim \frac{\Delta}{1+(\lambda/\sigma^2)} \exp\left\{-\frac{\Delta^2}{8} + \frac{\Delta^2}{8(1+\lambda/\sigma^2)}\right\}.$$

PROOF. See Appendix A.1.2. □

REMARK 2. The upper bound of Lemma 2 reveals that $\lambda/\sigma^2 \to \infty$ implies $R_x^* - R_z^* \to 0$ irrespective of the magnitude of Δ. Regarding to Δ, we also find that $R_x^* - R_z^* \to 0$ in the following scenarios: (1) if $\Delta \to 0$, irrespective of λ/σ^2, (2) if $\Delta \to \infty$ and $\lambda/\sigma^2 \not\to 0$, (3) if $\Delta \asymp 1$ and $\lambda/\sigma^2 \to \infty$.

The lower bound of Lemma 2, on the other hand, establishes the irreducible error for not observing Z. This term will naturally appear in the minimax lower bounds of the excess risk derived in the next section.

2.2. *Minimax lower bounds of the excess risk.* In this section, we establish minimax lower bounds of the excess risk $R_x(\widehat{g}) - R_z^*$ under model (1.1) and (1.3) for any classifier $\widehat{g}$. Here,

$$R_x(\widehat{g}) := \mathbb{P}\big\{\widehat{g}(X) \neq Y | \boldsymbol{D}\big\} \tag{2.4}$$

is the (conditional) misclassification error, given the training data

$$\boldsymbol{D} := (\boldsymbol{X}, \boldsymbol{Y}) = \big\{(X_1, Y_1), \ldots (X_n, Y_n)\big\}.$$

The results are established over the parameter space $\Theta(\lambda, \sigma, \Delta)$ in (2.3), which is characterized by three quantities: λ, σ^2 and Δ, all of which are allowed to grow with the sample size n. Our minimax lower bounds of the excess risk fully characterize the dependence on these quantities, in addition to the dimensions K and p and the sample size n.

We use $\mathbb{P}_\theta^{\boldsymbol{D}}$ to denote the set of all distributions of $\boldsymbol{D}$ parametrized by $\theta \in \Theta(\lambda, \sigma, \Delta)$ under models (1.1) and (1.3). For simplicity, we drop the dependence on θ for both $R_x(\widehat{g})$ and R_z^*. Define

$$\omega_n^* = \sqrt{\frac{K}{n} + \frac{\sigma^2}{\lambda}\Delta^2 + \frac{\sigma^2}{\lambda}\frac{\sigma^2 p}{\lambda n}\Delta^2}. \tag{2.5}$$

The following theorem states the minimax lower bounds of the excess risk for any classifier over the parameter space $\Theta(\lambda, \sigma, \Delta)$.

THEOREM 3. *Under model* (1.1), *assume* (i)–(iv), $K \geq 2$, $K/(n \wedge p) \leq c_1$, $\sigma^2/\lambda \leq c_2$ *and* $\sigma^2 p/(\lambda n) \leq c_3$ *for some sufficiently small constants* $c_1, c_2, c_3 > 0$. *There exists some constants* $c_0 \in (0, 1)$ *and* $C > 0$ *such that*:

1. *If* $\Delta \asymp 1$, *then*
$$\inf_{\widehat{g}} \sup_{\theta \in \Theta(\lambda,\sigma,\Delta)} \mathbb{P}_\theta^{\boldsymbol{D}} \{R_x(\widehat{g}) - R_z^* \geq C(\omega_n^*)^2\} \geq c_0.$$
2. *If* $\Delta \to \infty$ *and* $\sigma^2/\lambda = o(1)$ *as* $n \to \infty$, *then*
$$\inf_{\widehat{g}} \sup_{\theta \in \Theta(\lambda,\sigma,\Delta)} \mathbb{P}_\theta^{\boldsymbol{D}} \left\{ R_x(\widehat{g}) - R_z^* \geq C(\omega_n^*)^2 \exp\left\{ -\left[\frac{1}{8} + o(1)\right]\Delta^2 \right\} \right\} \geq c_0.$$
3. *If* $\Delta \to 0$ *as* $n \to \infty$, *then*
$$\inf_{\widehat{g}} \sup_{\theta \in \Theta(\lambda,\sigma,\Delta)} \mathbb{P}_\theta^{\boldsymbol{D}} \left\{ R_x(\widehat{g}) - R_z^* \geq C \min\left\{ \frac{\omega_n^*}{\Delta}, 1 \right\} \omega_n^* \right\} \geq c_0.$$

The infima in all statements are taken over all classifiers.

PROOF. The proof of Theorem 3 is deferred to Appendix B. □

The lower bounds in Theorem 3 consist of three terms: the one related with K/n is the optimal rate of the excess risk even when Z were observable; the second one related with σ^2/λ is the irreducible error for not observing Z (see, Lemma 1); the last one involving $\sigma^2 p/(\lambda n)$ is the price to pay for estimating the column space of A. Although the third term could get absorbed by the second term as $\sigma^2 p/(\lambda n) \leq c_3$, we incorporate it here for transparent interpretation. The lower bounds in Theorem 3 are tight as we show in Section 5.3 that there exists a classifier whose excess risk has a matching upper bound.

REMARK 3 (Phase transition in Δ). Recall from (2.2) that Δ quantifies the separation between $N(\alpha_0, \Sigma_{Z|Y})$ and $N(\alpha_1, \Sigma_{Z|Y})$. We see in Theorem 3 a phase transition of the rates of convergence of the excess risk as Δ varies. When Δ is of constant order, the excess risk has minimax convergence rate
$$\frac{K}{n} + \frac{\sigma^2}{\lambda} + \frac{\sigma^2}{\lambda}\frac{\sigma^2 p}{\lambda n}.$$
When $\Delta \to \infty$, we see that the minimax rate of convergence of the excess risk gets faster exponentially in Δ^2. For instance, if $\Delta^2 \geq C_0 \log n$ for some constant $C_0 > 0$, then the minimax rate already becomes *polynomially faster in n* as
$$\left[\frac{K}{n} + \frac{\sigma^2}{\lambda} + \frac{\sigma^2}{\lambda}\frac{\sigma^2 p}{\lambda n}\right]\frac{1}{n^{C_1}}$$
for some $C_1 > 0$ depending on C_0. The condition $\sigma^2/\lambda = o(1)$ for $\Delta \to \infty$ can be removed, and the lower bound remains the same except the factor $(1/8)$ gets replaced by $(1/8)(1/(1 + \lambda/\sigma^2))$. Finally, when $\Delta \to 0$, a more challenging, yet important case, the minimax convergence rate of the excess risk gets slower. It is worth noting that although the oracle Bayes risk $R_z^* \to 1/2$ when $\Delta \to 0$, the minimax excess risk still converges to zero at least in ω_n^*-rate. If $\omega_n^* \lesssim \Delta$, the convergence gets faster as
$$\frac{K}{n}\frac{1}{\Delta} + \frac{\sigma^2}{\lambda}\Delta + \frac{\sigma^2}{\lambda}\frac{\sigma^2 p}{\lambda n}\Delta.$$

REMARK 4 (Proof technique). To prove Theorem 3, the three terms in the lower bound are derived separately in the setting where $X|Y$ is Gaussian. Since, for any classifier $\widehat{g}$,

$$R_x(\widehat{g}) - R_z^* = \big(R_x(\widehat{g}) - R_x^*\big) + \big(R_x^* - R_z^*\big),$$

in view of Lemma 1, it suffices to prove the two terms related with K/n and $\sigma^2 p/(\lambda n)$ constitute the lower bounds of $R_x(\widehat{g}) - R_x^*$. In fact, as a byproduct of our result, we also derive minimax lower bounds of the excess risk relative to R_x^*. This derivation is based on constructing subsets of $\Theta(\lambda, \sigma, \Delta)$ by fixing either A or α_0 and α_1 separately. The choice of A is based on the hypercube construction for matrices with orthonormal columns [50], Lemma A.5. The analyses of both terms are nonstandard as the excess risk is not a semidistance, as required by standard techniques of proving minimax lower bounds. Based on a reduction scheme established in Appendix B, we show that proving Theorem 3 suffices to establish a minimax lower bound of the following loss function:

$$L_\theta(\widehat{g}) := \mathbb{P}_\theta\big\{\widehat{g}(X) \neq g_\theta^*(X) | \boldsymbol{D}\big\}.$$

Here, $\mathbb{P}_\theta$ is taken with respect to X and $g_\theta^*(X)$ is the Bayes rule based on X that minimizes $R_x(g)$ over $g : \mathbb{R}^p \to \{0, 1\}$. Since $L_\theta(\widehat{g})$ is shown to satisfy a local triangle inequality-type bound such that a variant of Fano's lemma can be applied [4], Proposition 2, we proved a crucial result, in Lemmas B.5 and B.6 of Appendix B, that

$$\inf_{\widehat{g}} \sup_{\theta \in \Theta(\lambda,\sigma,\Delta)} \mathbb{P}_\theta^{\boldsymbol{D}} \left\{ L_\theta(\widehat{g}) \geq C\left(\sqrt{\frac{K}{n}}\frac{1}{\Delta} + \sqrt{\frac{\sigma^2}{\lambda}\frac{\sigma^2 p}{\lambda n}}\right) e^{-\frac{\Delta^2}{8}} \right\} \geq c_0 \tag{2.6}$$

for some constant $c_0 \in (0, 1)$ and $C > 0$.

REMARK 5 (Comparison with the existing literature). As mentioned above, a byproduct of our proof of Theorem 3 is the minimax lower bounds of $R_x(\widehat{g}) - R_x^*$ in the setting where $X|Y$ is Gaussian, which have exactly the same form as Theorem 3 but without the second term related with σ^2/λ. It is informative to put this lower bound of $R_x(\widehat{g}) - R_x^*$ in comparison to the existing literature in this special setting.

Under the classical LDA model (1.5), [20] derives the minimax lower bounds of $R_x(\widehat{g}) - R_x^*$ over a suitable parameter space for $\Delta \gtrsim 1$, which have the same form as ours with $K/n + \sigma^4 p\Delta^2/(\lambda^2 n)$ replaced by s/n for $s := \|\Sigma^{-1}(\mu_1 - \mu_0)\|_0$. In contrast, our lower bounds reflect the benefit of considering an approximate lower-dimensional structure of $X|Y$ under (1.1) and (1.5) instead of directly assuming sparsity on $\Sigma^{-1}(\mu_1 - \mu_0)$. These two lower bounds coincide in the low-dimensional setting ($p < n$) when there is no sparsity in $\Sigma^{-1}(\mu_1 - \mu_0)$, that is, $s = p$, and when there is no low-dimensional hidden factor model (i.e., $X = Z$ with $K = p$, $A = \boldsymbol{I}_p$ and $W = \boldsymbol{0}_p$). On the other hand, [19] only established the phase transition between $\Delta \asymp 1$ and $\Delta \to \infty$ whereas we are able to derive the minimax lower bound for $\Delta \to 0$, a case that has not even been analyzed in the classical LDA literature.

Technically, it is also worth mentioning that the latent model structure on X via (1.1) brings considerable additional difficulties for establishing the lower bounds of $R_x(\widehat{g}) - R_x^*$. Indeed, for any $\theta \in \Theta(\lambda, \sigma, \Delta)$, the covariance matrix of $X|Y$ is $\Sigma(\theta) = A\Sigma_{Z|Y}A^\top + \Sigma_W$, which cannot be chosen as a diagonal matrix to simplify the analysis as done by [20]. Furthermore, to derive the term $\sigma^4 p\Delta^2/(\lambda^2 n)$ in the lower bound for quantifying the error of estimating the column space of A, we need to carefully choose the subset of $\Theta(\lambda, \sigma, \Delta)$ via the hypercube construction ([50], Lemma A.5) that has been used for proving the optimal rates of the sparse PCA. Since the statistical distance (such as KL-divergence) between any two of thus constructed hypotheses of $\Theta(\lambda, \sigma, \Delta)$ is diverging whenever $p/n \to \infty$ (see Lemma B.4 in Appendix B), a different analysis than the standard one (for instance, in [4]) has to be used to allow $p > n$ and a large amount of work is devoted to provide a meaningful and sharp lower bound that is valid for both $p < n$ and $p > n$ (see Lemma B.5 for details).

3. Methodology. In this section, we describe our classification method based on n i.i.d. observations from models (1.1) and (1.3). We first state a general method in Section 3.1 which is motivated by the optimal oracle rule g_z^* in (1.6) and (1.7), and is based on prediction of the unobserved factors $Z_1, \ldots, Z_n, Z$ in the features $X_1, \ldots, X_n, X$ by projections. In Section 3.2, we state our proposed methods via principal component projections as well as a cross-fitting strategy for high-dimensional scenarios. Selection of the number of principal components is further discussed in Section 3.3.

3.1. *General approach.* The first idea is to change the classification problem into a regression problem, at the population level. The close relationship between LDA and regression has been observed before; see, for instance, Section 8.3.3 in [33, 35] and [40]. Let $\Sigma_Z = \mathrm{Cov}(Z)$ be the unconditional covariance matrix of Z. Define

$$\begin{aligned} \beta &= \pi_0\pi_1 \Sigma_Z^{-1}(\alpha_1 - \alpha_0), \\ \beta_0 &= -\frac{1}{2}(\alpha_0 + \alpha_1)^\top \beta + \pi_0\pi_1\big[1 - (\alpha_1 - \alpha_0)^\top \beta\big] \log \frac{\pi_1}{\pi_0}. \end{aligned} \tag{3.1}$$

PROPOSITION 4. *Let η, η_0 and β, β_0 be defined in* (1.7) *and* (3.1)*, respectively. Under model* (1.3) *and assumption* (iv)*, we have*

$$z^\top \eta + \eta_0 \geq 0 \quad \Longleftrightarrow \quad z^\top \beta + \beta_0 \geq 0.$$

Furthermore,

$$\beta = \Sigma_Z^{-1} \mathrm{Cov}(Z, Y).$$

PROOF. The proof of Proposition 4 can be found in Appendix A.2. □

REMARK 6. In fact, our proof shows that the first statement of Proposition 4 still holds if we replace $\pi_0\pi_1$ in the definition of β by any positive value coupled with corresponding modification of β_0 (see Lemma A.1 in Appendix A.2 for the precise statement). The advantage of using $\pi_0\pi_1$ in (3.1) is that β can be obtained by simply regressing Y on Z. For this choice of β, our proof also reveals

$$z^\top \eta + \eta_0 = \frac{1}{\pi_0\pi_1[1 - (\alpha_1 - \alpha_0)^\top \beta]}(z^\top \beta + \beta_0) = \frac{1 + \pi_0\pi_1\Delta^2}{\pi_0\pi_1}(z^\top \beta + \beta_0), \tag{3.2}$$

a key identity that will used later in Section 8 to extend our approach for handling multiclass classification problems.

Proposition 4 implies the equivalence between the linear rules $g_z^*(z)$ in (1.7) and

$$g_z(z) := \mathbb{1}\{z^\top \beta + \beta_0 \geq 0\} \tag{3.3}$$

based on, respectively, the half-spaces $\{z | z^\top \eta + \eta_0 \geq 0\}$ and $\{z | z^\top \beta + \beta_0 \geq 0\}$. According to Proposition 4, if $\boldsymbol{Z} = (Z_1^\top, \ldots, Z_n^\top)^\top \in \mathbb{R}^{n \times K}$ were observed, it is natural to use the least squares estimator $(\boldsymbol{Z}^\top \Pi_n \boldsymbol{Z})^+ \boldsymbol{Z}^\top \Pi_n \boldsymbol{Y}$ to estimate β. Recall that $\Pi_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix and M^+ is the Moore–Penrose inverse of any matrix M. Since in practice only $\boldsymbol{X} = (X_1^\top, \ldots, X_n^\top)^\top \in \mathbb{R}^{n \times p}$ is observed, we propose to estimate $z^\top \beta$ by

$$x^\top \widehat{\theta} := x^\top B(\Pi_n \boldsymbol{X} B)^+ \boldsymbol{Y} = x^\top B\big(B^\top \boldsymbol{X}^\top \Pi_n X B\big)^+ B^\top \boldsymbol{X}^\top \Pi_n \boldsymbol{Y} \tag{3.4}$$

with $x \in \mathbb{R}^p$ being one realization of X from model (1.1). Here, in principal $B \in \mathbb{R}^{p\times q}$ could be any matrix with any $q \in \{1, \ldots, p\}$. Furthermore, we estimate β_0 by

$$\widehat{\beta}_0 := -\frac{1}{2}(\widehat{\mu}_0 + \widehat{\mu}_1)^\top \widehat{\theta} + \widehat{\pi}_0\widehat{\pi}_1\big[1 - (\widehat{\mu}_1 - \widehat{\mu}_0)^\top \widehat{\theta}\big]\log\frac{\widehat{\pi}_1}{\widehat{\pi}_0} \tag{3.5}$$

based on standard nonparametric estimates

$$n_k = \sum_{i=1}^n \mathbb{1}\{Y_i = k\}, \qquad \widehat{\pi}_k = \frac{n_k}{n}, \qquad \widehat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^n X_i\mathbb{1}\{Y_i = k\}, \quad k \in \{0, 1\}. \tag{3.6}$$

Our final classifier is

$$\widehat{g}_x(x) := \mathbb{1}\{x^\top \widehat{\theta} + \widehat{\beta}_0 \geq 0\}. \tag{3.7}$$

Notice that $\widehat{\theta}$, $\widehat{\beta}_0$ and $\widehat{g}_x(x)$ all depend on B implicitly.

3.2. *Principal component (PC) based classifiers.* Though the classifier in (3.7) can use any matrix B, in this paper we mainly consider the choice $B = \boldsymbol{U}_r \in \mathbb{R}^{p\times r}$, for some $r \in \{1, \ldots, p\}$, where the matrix $\boldsymbol{U}_r$ consists of the first r right-singular vectors of $\Pi_n\boldsymbol{X}$, the centered $\boldsymbol{X}$. In this case, $x^\top\widehat{\theta}$ is the famous principal component regression (PCR) predictor by using r principal components [34]. The optimal choice of r would be K, the number of latent factors when it is known in advance. We analyze the classifier with $B = \boldsymbol{U}_K$ in Theorem 9 of Section 5.2.

Suggested by our theory, in the high-dimensional setting $p > n$, performance of the PC-based classifiers can be improved either by using an additional data set or via data-splitting.

In several applications, such as semisupervised learning, researchers also have access to an additional set of unlabeled data. Given an additional data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n'\times p}$ with i.i.d. (unlabeled) observations from model (1.1) with $n' \asymp n$ and independent of $\boldsymbol{X}$ in (3.4), it is often beneficial to use $B = \widetilde{\boldsymbol{U}}_K$ based on the first K right singular vectors of $\Pi_{n'}\tilde{\boldsymbol{X}}$. This classifier is analyzed in Theorem 10 of Section 5.2.

When additional data is not available, we advocate to use a sample splitting technique called k-fold cross-fitting [21]. First, we randomly split the data into k folds, and for each fold, we use it as $\tilde{\boldsymbol{X}}$ to construct $\widetilde{\boldsymbol{U}}_r$ and use the remaining data as $\boldsymbol{X}$ and $\boldsymbol{Y}$ to obtain $\widehat{\theta}$ and $\widehat{\beta}_0$ from (3.4) and (3.5), respectively. In the end, the final classifier is constructed via (3.7) based on the averaged k pairs of $\widehat{\theta}$ and $\widehat{\beta}_0$. Theoretically, it is straightforward to show that the resulting classifiers share the same conclusions as Theorem 10 for $k = \mathcal{O}(1)$. Empirically, since this cross-fitting strategy ultimately uses all data points, it might mitigate the efficiency loss due to sample splitting. Standard choices of k include $k = 2$ and $k = 5$ while the latter is reported to have smaller standard errors [21].

3.3. *Estimation of the number of retained PCs.* When K is unknown, we propose to estimate it by

$$\widehat{K} := \operatorname*{arg\,min}_{k\in\{0,1,\ldots,\bar{K}\}} \frac{\sum_{j>k}\sigma_j^2}{np - c_0(n+p)k}, \quad \text{with } \bar{K} := \left\lfloor \frac{\nu}{2c_0(1+\nu)}(n \wedge p) \right\rfloor, \tag{3.8}$$

for absolute constants c_0 and $\nu > 1$. The latter is introduced to avoid division by zero and can be set arbitrarily large. The choice of $c_0 = 2.1$ is used in all of our simulations and has overall good performance. The sum $\sum_j \sigma_j u_j v_j^\top$, with nonincreasing σ_j, is the singular-value-decomposition (SVD) of $\Pi_n\boldsymbol{X}$ or $\Pi_n\tilde{\boldsymbol{X}}$.

Criterion (3.8) was originally proposed in [16] for selecting the rank of the coefficient of a multivariate response regression model and is further adopted by [13] for selecting the

number of retained principal components under the framework of factor regression models. It also has close connection to the well-known elbow method, but is more practical in terms of parameter tuning. The main computation of solving (3.8) is to compute the SVD of $\Pi_n \boldsymbol{X}$ once. In Section 5.1, we show the consistency of $\widehat{K}$, ensuring that the classifier with $B = \boldsymbol{U}_{\widehat{K}}$ shares the same theoretical properties as the one with $B = \boldsymbol{U}_K$.

4. A general strategy of bounding the excess classification error. In this section, we establish a general theory for analyzing the excess risk of the classifier $\widehat{g}_x$ in (3.7) that uses *any* matrix B for the estimate $\widehat{\theta}$ in (3.4). The main purpose is to establish high-level conditions that yield a consistent classifier constructed in Section 3 in the sense

$$R_x(\widehat{g}_x) := \mathbb{P}\{\widehat{g}_x(X) \neq Y | \boldsymbol{D}\} \to R_z^*, \quad \text{in probability, as } n \to \infty$$

and further to provide its rate of convergence. We recall that $\mathbb{P}$ is taken with respect to (X, Y).

For convenience, we introduce the notation

$$\widehat{G}_x(x) := x^\top \widehat{\theta} + \widehat{\beta}_0, \qquad G_z(z) := z^\top \beta + \beta_0 \tag{4.1}$$

such that $\widehat{g}_x(x) = \mathbb{1}\{\widehat{G}_x(x) \geq 0\}$ from (3.7) and, using the equivalence in Proposition 4,

$$g_z^*(z) = \mathbb{1}\{G_z(z) \geq 0\}. \tag{4.2}$$

Recall that $\widehat{g}_x$ depends on the choice of B via $\widehat{\theta}$ and $\widehat{\beta}_0$.

The following theorem provides a general bound for the excess risk of $\widehat{g}_x$ that uses any B in (3.4). Its proof can be found in Appendix A.3.1.

THEOREM 5. *Under model* (1.1), *assume* (i)–(iv). *For all* $t > 0$, *we have*

$$R_x(\widehat{g}_x) - R_z^* \leq \mathbb{P}\{|\widehat{G}_x(X) - G_z(Z)| > t | \boldsymbol{D}\} + c_* t P(t), \tag{4.3}$$

where $c_* = \Delta^2 + (\pi_0 \pi_1)^{-1}$ *and*

$$P(t) = \pi_0[\Phi(R) - \Phi(R - t c_*/\Delta)] + \pi_1[\Phi(L + t c_*/\Delta) - \Phi(L)] \tag{4.4}$$

with

$$L = -\frac{\Delta}{2} - \frac{\log \frac{\pi_1}{\pi_0}}{\Delta}, \qquad R = \frac{\Delta}{2} - \frac{\log \frac{\pi_1}{\pi_0}}{\Delta}.$$

REMARK 7. The quantity $P(t)$ in (4.4) is in fact

$$\pi_0 \mathbb{P}\{-t < G_z(Z) < 0 | Y = 0\} + \pi_1 \mathbb{P}\{0 < G_z(Z) < t | Y = 1\},$$

which describes the probabilistic behavior of the margin of the hyperplane $\{z : G_z(z) = 0\}$ that separates the distributions $Z|Y = 0$ and $Z|Y = 1$. Conditions that control the margin between $Z|Y = 0$ and $Z|Y = 1$ are more suitable in our current setting and have a different perspective than the usual margin condition in [49] that controls the probability $\mathbb{P}\{|\eta(Z) - 1/2| < \delta\}$ for any $0 \leq \delta \leq 1/2$, with $\eta(z) := \mathbb{P}(Y = 1 | Z = z)$.

REMARK 8 (Extension to nonlinear classifiers). The proof of Theorem 5 also allows us to analyze more complex classifiers. Indeed, let $\Lambda_z(z)$ be the logarithm of the ratio between $\mathbb{P}(Z = z, Y = 1)$ and $\mathbb{P}(Z = z, Y = 0)$, and let $\widehat{\Lambda}_x(x)$ be an arbitrary estimate of $\Lambda_z(z)$. We can easily derive from our proof of Theorem 5 the following excess risk bound for the classifier $\widehat{g}_x(x) = \mathbb{1}\{\widehat{\Lambda}_x(x) \geq 0\}$:

$$\begin{aligned} R_x(\widehat{g}_x) - R_z^* \leq{}& \mathbb{P}\{|\widehat{\Lambda}_x(X) - \Lambda_z(Z)| > t | \boldsymbol{D}\} \\ &+ t\pi_0 \mathbb{P}\{-t < \Lambda_z(Z) < 0 | Y = 0\} + t\pi_1 \mathbb{P}\{0 < \Lambda_z(Z) < t | Y = 1\}, \end{aligned} \tag{4.5}$$

for any $t > 0$. Therefore, bound in (4.5) can be used as an initial step for analyzing any classification problems, particularly suitable for situations where conditional distributions $Z|Y$ are specified. The remaining difficulty is to find a good estimator $\widehat{\Lambda}_x(x)$ and to control $|\widehat{\Lambda}_x(X) - \Lambda_z(Z)|$. For instance, when $Z|Y = k$, for $k \in \{0, 1\}$, have Gaussian distributions with different means and different covariances, the Bayes rule of using Z (equivalently, $\Lambda_z(Z)$) becomes quadratic, leading to an estimator $\widehat{\Lambda}_x(x)$ that is quadratic in x as well. Since both the procedure and the analysis are different, we will study this setting in a separate paper.

From (4.1), we find the identity

$$\widehat{G}_x(X) - G_z(Z) = Z^\top\big(A^\top\widehat{\theta} - \beta\big) + W^\top\widehat{\theta} + \widehat{\beta}_0 - \beta_0. \tag{4.6}$$

To establish its deviation inequalities, our analysis uses the following distributional assumption on W from (1.1). We assume that:

(v) $W = \Sigma_W^{1/2}\widetilde{W}$ and $\widetilde{W}$ is a mean-zero γ-sub-Gaussian random vector with $\mathbb{E}[\widetilde{W}\widetilde{W}^\top] = \boldsymbol{I}_p$ and $\mathbb{E}[\exp(u^\top\widetilde{W})] \le \exp(\gamma^2/2)$, for all $\|u\|_2 = 1$.

We stress that the distributions of $X|Y$ need not be Gaussian. In addition, we require that

(vi) π_0 and π_1 are fixed and bounded from below by some constant $c \in (0, 1/2]$.

The following proposition states a deviation inequality of $|\widehat{G}_x(X) - G_z(Z)|$, which holds with high probability under the law $\mathbb{P}^{\boldsymbol{D}}$. It depends on three quantities:

$$\widehat{r}_1 := \big\|\Sigma_Z^{1/2}\big(A^\top\widehat{\theta} - \beta\big)\big\|_2, \qquad \widehat{r}_2 := \|\widehat{\theta}\|_2, \qquad \widehat{r}_3 := \frac{1}{\sqrt{n}}\big\|\boldsymbol{W}(P_B - P_A)\big\|_{\mathrm{op}}. \tag{4.7}$$

For any matrix M, let P_M denote the projection onto its column space. From (4.6), appearance of the first two quantities in (4.7) is natural since Z and W are independent of $\widehat{\theta}$ and $\widehat{\beta}_0$, and Z and W are sub-Gaussian random vectors under the distributional assumptions (iv) and (v). The third quantity $\|\boldsymbol{W}(P_B - P_A)\|_{\mathrm{op}}$ in (4.7) originates from $\widehat{\beta}_0 - \beta_0$ and reflects the benefit of using a matrix B that estimates the column space of A well.

PROPOSITION 6. *Under model* (1.1), *assume* (i)–(vi) *and* $K \log n \le cn$ *for some constant* $c > 0$. *For any* $a \ge 1$, *we have*

$$\mathbb{P}^{\boldsymbol{D}}\big\{\big|\widehat{G}_x(X) - G_z(Z)\big| \ge \widehat{\omega}_n(a)|\boldsymbol{D}\big\} \lesssim n^{-a}\big\} = 1 - \mathcal{O}\big(n^{-1}\big). \tag{4.8}$$

Here, *for some constant* $C > 0$ *depending on* γ *only*,

$$\widehat{\omega}_n(a) = C\left\{\sqrt{a\log n}\big(\widehat{r}_1 + \|\Sigma_W\|_{\mathrm{op}}^{1/2}\widehat{r}_2\big) + \widehat{r}_2\widehat{r}_3 + \sqrt{\frac{\log n}{n}}\right\}. \tag{4.9}$$

PROOF. See Appendix A.3.2. □

Proposition 6 implies that we need to control $\widehat{\omega}_n(a)$ whose randomness solely depends on $\boldsymbol{D}$. In view of Theorem 5 and Proposition 6, we have the following result.

THEOREM 7. *Under model* (1.1), *assume* (i)–(vi) *and* $K \log n \le cn$ *for some constant* $c > 0$. *For any* $a \ge 1$ *and any sequence* $\omega_n > 0$, *on the event* $\{\widehat{\omega}_n(a) \le \omega_n\}$, *the following holds with probability* $1 - \mathcal{O}(n^{-1})$ *under the law* $\mathbb{P}^{\boldsymbol{D}}$:

$$R_x(\widehat{g}_x) - R_z^* \lesssim n^{-a} + \begin{cases} \omega_n^2 & \text{if } \Delta \asymp 1, \\ \omega_n^2\exp\{-[c_\pi + o(1)]\Delta^2\} & \text{if } \Delta \to \infty \text{ and } \omega_n = o(1), \\ \omega_n^2\exp\{-[c' + o(1)]/\Delta^2\} & \text{if } \Delta \to 0,\ \pi_0 \ne \pi_1 \text{ and } \omega_n = o(1), \\ \omega_n\min\{1, \omega_n/\Delta\} & \text{if } \Delta \to 0 \text{ and } \pi_0 = \pi_1. \end{cases}$$

Here, c_π *and* c' *are some absolute positive constants and* $c_\pi = 1/8$ *if* $\pi_0 = \pi_1$.

Hence, it remains to find a deterministic sequence $\omega_n \to 0$ such that $\mathbb{P}^{\boldsymbol{D}}\{\widehat{\omega}_n(a) \le \omega_n\} \to 1$ as $n \to \infty$. Further, in view of (4.9), all we need is to find deterministic upper bounds of $\widehat{r}_1, \widehat{r}_2$ and $\widehat{r}_3$. In such way, Theorem 7 serves as a general tool for analyzing the excess risk of the classifier constructed via (3.4)–(3.7) by using any matrix B.

Later in Section 5, we apply Theorem 7 to analyze several classifiers, including the principal components based classifier by choosing $B = \boldsymbol{U}_K$ and $B = \widetilde{\boldsymbol{U}}_K$ as well as their counterparts based on the data-dependent choice $\widehat{K}$. For these PC-based classifiers, we will find a sequence ω_n that closely matches the sequence ω_n^* in (2.5) under suitable conditions, up to $\log(n)$, for our procedure. In view of Theorem 3, this rate turns out to be minimax-optimal over a subset of the parameter space considered in Theorem 3, up to $\log(n)$ factors.

Although not pursued in this paper, it is worth mentioning some other reasonable choices of B including, for instance, the identity matrix $\boldsymbol{I}_p$, which leads to the generalized least squares based classifier [14], the estimator of A in [12], the projection matrix from supervised PCA [7, 9] and the projection matrix obtained via partial least squares regression [8, 42].

REMARK 9. We observe the same phase transition in Theorem 7 for $\Delta \asymp 1$ and $\Delta \to \infty$ as discussed in Remark 3. To the best of our knowledge, upper bounds of the excess risk in the regime $\Delta = o(1)$ are not known in the existing literature. Our result in this regime relies on a careful analysis, which does not require any condition on Δ, in contrast to the existing analysis of the classical high-dimensional LDA problems. For instance, under model (1.5), [19] assumes $\Delta_x^2 := (\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 - \mu_0) \gtrsim 1$ and $\Delta_x^2(s \log n/n) = o(1)$ to derive the convergence rate of their estimator of $\Sigma^{-1}(\mu_1 - \mu_0)$ with $s = \|\Sigma^{-1}(\mu_1 - \mu_0)\|_0$. As a result, their results of excess misclassification risk only hold for $\Delta_x \gtrsim 1$.

5. Rates of convergence of the PC-based classifier. We apply our general theory in Section 4 to several classifiers corresponding to different choices of $B = \boldsymbol{U}_K$, $B = \boldsymbol{U}_{\widehat{K}}$, $B = \widetilde{\boldsymbol{U}}_K$ and $B = \widetilde{\boldsymbol{U}}_{\widehat{K}}$ in (3.4). Since our analysis is beyond the parameter space $\Theta(\lambda, \sigma, \Delta)$ in (2.3), we first generalize the signal-to-noise ratio λ/σ^2 of predicting Z from X given Y by introducing

$$\xi^* := \frac{\lambda_K(A\Sigma_{Z|Y}A^\top)}{\lambda_1(\Sigma_W)}. \tag{5.1}$$

We also need the related quantity

$$\xi := \frac{\lambda_K(A\Sigma_{Z|Y}A^\top)}{\delta_W}, \tag{5.2}$$

that characterizes the signal-to-noise ratio of predicting $\boldsymbol{Z}$ from $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$. Indeed, note that we replaced $\lambda_1(\Sigma_W)$ in (5.1) by

$$\delta_W = \lambda_1(\Sigma_W) + \frac{\operatorname{tr}(\Sigma_W)}{n} \tag{5.3}$$

and the largest eigenvalue of the random matrix $\boldsymbol{W}^\top \boldsymbol{W}/n$ is of order $\mathcal{O}_{\mathbb{P}}(\delta_W)$ under assumption (v) (see, for instance, [13], Lemma 22).

5.1. *Consistent estimation of the latent dimension K.* Since in practice the true K is often unknown, we analyze the estimated rank $\widehat{K}$ selected from (3.8).

Consistency of $\widehat{K}$ under the factor model (1.1) when Z is a zero-mean sub-Gaussian random vector has been established in [13], Proposition 8. Here, we establish such property of $\widehat{K}$ under (1.1) where Z follows a mixture of two Gaussian distributions. Let $r_e(\Sigma_W) = \operatorname{tr}(\Sigma_W)/\lambda_1(\Sigma_W)$ denote the effective rank of Σ_W.

THEOREM 8. *Let $\widehat{K}$ be defined in (3.8) for some absolute constant $c_0 > 0$. Under model (1.1), assume* (i)–(vi), *and, in addition,*

$$K \le \bar{K}, \xi \ge C \quad \text{and} \quad r_e(\Sigma_W) \ge C'(n \wedge p)$$

for some constants $C, C' > 0$. Then

$$\mathbb{P}^{\boldsymbol{D}}\{\widehat{K} = K\} = 1 - \mathcal{O}(n^{-1}).$$

PROOF. The proof is deferred to Appendix A.4.1. □

Theorem 8 implies that the classifier that uses $B = \boldsymbol{U}_{\widehat{K}}$ ($B = \widetilde{\boldsymbol{U}}_{\widehat{K}}$) has the same excess risk bound as that uses $B = \boldsymbol{U}_K$ ($B = \widetilde{\boldsymbol{U}}_K$). For this reason, we restrict our analysis in the remaining of this section to B based on the first K principal components of $\boldsymbol{U}$ and $\widetilde{\boldsymbol{U}}$.

The condition $K \le \bar{K}$ holds, for instance, if $K \le c'(n \wedge p)$ with $c' \le \nu/(2c_0(1+\nu))$. Condition $r_e(\Sigma_W) \ge C'(n \wedge p)$ holds, for instance, in the commonly considered setting

$$0 < c \le \lambda_p(\Sigma_W) \le \lambda_1(\Sigma_W) \le C < \infty$$

while being more general.

The condition that $\xi \ge C$ is also needed in our subsequent derivation of the rates of the excess risks for the classifiers using $B = \boldsymbol{U}_K$ and $B = \widetilde{\boldsymbol{U}}_K$. This essentially requires $\xi^* \ge C$ in the low-dimensional settings, and $\xi^* \ge C(p/n)$ in the high-dimensional settings (see Remark 12 below for details). Since the minimax lower bounds for the excess risk in Theorem 3 above contain the term $\min(1, \Delta)/\xi^*$, it is imperative that the signal-to-noise ratio ξ^* is large to guarantee good performance of the classifier, irrespective of the estimation of the latent dimension K.

We investigate in Appendix E.1 the consequences of inconsistent estimates $\widehat{K}$ and found that our proposed classifiers are robust against both underestimation and overestimation. This is corroborated in our follow-up work [14], that proves that the classifier using $\widehat{\theta} = (\Pi_n \boldsymbol{X})^+ \boldsymbol{Y}$ based on $B = \boldsymbol{I}_p$ (in other words, $\widehat{K} = p$), often is minimax optimal and performing slightly inferior to $B = \boldsymbol{U}_K$ in finite sample simulations.

5.2. *PC-based LDA by using the true dimension K.* The following theorem states the excess risk bounds of $\widehat{g}_x$ that uses $B = \boldsymbol{U}_K$. Its proof can be found in Appendix A.4.2. Denote by κ the condition number $\lambda_1(A\Sigma_Z A^\top)/\lambda_K(A\Sigma_Z A^\top)$ of the matrix $A\Sigma_Z A^\top$.

THEOREM 9. *Under model (1.1), assume* (i)–(vi). *If $K \log n \le cn$ and $\xi \ge C\kappa^2$ for some constants $c, C > 0$, then for any $a \ge 1$ and*

$$\omega_n(a) = \left(\sqrt{\frac{K \log n}{n}} + \min\{1, \Delta\}\sqrt{\frac{1}{\xi^*}} + \sqrt{\frac{\kappa}{\xi^2}}\right)\sqrt{a \log n}, \tag{5.4}$$

we have $\mathbb{P}^{\boldsymbol{D}}\{\widehat{\omega}_n(a) \lesssim \omega_n(a)\} = 1 - \mathcal{O}(n^{-1})$. Hence, with this probability, the conclusion of Theorem 7 holds for the classifier that uses $B = \boldsymbol{U}_K$ for $\omega_n(a)$ in (5.4).

Theorem 9 requires $\xi \ge C\kappa^2$, which can be relaxed to $\xi \ge C$, as shown in the proof (see Remark 1 in Appendix A.4). However, the stronger condition can lead to a faster rate when one has additional data set to construct $B = \widetilde{\boldsymbol{U}}_K$, as stated in the theorem below. Its proof can be found in Appendix A.4.4.

THEOREM 10. *Under the same conditions of Theorem 9, for any $a > 0$ and*

$$\omega_n(a) = \left(\sqrt{\frac{K\log n}{n}} + \min\{1, \Delta\}\sqrt{\frac{1}{\xi^*}}\right)\sqrt{a\log n}, \tag{5.5}$$

we have $\mathbb{P}^{\boldsymbol{D}}\{\widehat{\omega}_n(a) \lesssim \omega_n(a)\} = 1 - \mathcal{O}(n^{-1})$. Hence, with this probability, the conclusion of Theorem 7 holds for the classifier that uses $B = \widetilde{\boldsymbol{U}}_K$ for $\omega_n(a)$ in (5.5).

REMARK 10 (Polynomially fast rates). In view of Theorems 9 and 10, fast rates (of the order $\mathcal{O}(n^{-a})$ for arbitrary $a \geq 1$) are obtained for both PC-based procedures, provided that (a) $\Delta^2 \gg \log n$ or (b) $1/\Delta^2 \gg \log n$ and $\pi_0 \neq \pi_1$.

REMARK 11 (Advantage of using an independent data set or data splitting). Compared to (5.4) in Theorem 9, the convergence rate of the excess risk of the classifier that uses $B = \widetilde{\boldsymbol{U}}_K$ does not have the third term $\sqrt{\kappa/\xi^2}$. This advantage only becomes evident when $p > n$ and ξ^* is not sufficiently large. We refer to Remark 12 below for detailed explanation.

To understand why using $\widetilde{\boldsymbol{U}}_K$, that is independent of $\boldsymbol{X}$, yields a smaller excess risk, recall that the third term in (5.4) originates from predicting $\boldsymbol{Z}$ from $\boldsymbol{X}$ and its derivation involves controlling $\|\boldsymbol{W}(P_{\boldsymbol{U}_K} - P_A)\|_{\mathrm{op}}$. Since $\boldsymbol{U}_K$ is constructed from $\boldsymbol{X}$, hence also depends on $\boldsymbol{W}$, the dependence between $\boldsymbol{W}$ and $\boldsymbol{U}_K$ renders a slow rate for $\|\boldsymbol{W}(P_{\boldsymbol{U}_K} - P_A)\|_{\mathrm{op}}$. The fact that auxiliary data can bring improvements (in terms of either smaller prediction/estimation error or weaker conditions) is a phenomenon that has been observed in other problems, such as the problem of estimating the optimal instrument in sparse high-dimensional instrumental variable model [10] and the problem of making inference on a low-dimensional parameter in the presence of high-dimensional nuisance parameters [21].

REMARK 12 (Simplified rates within $\Theta(\lambda, \sigma, \Delta)$). To obtain more insight from the results of Theorems 9 and 10, consider $\theta \in \Theta(\lambda, \sigma, \Delta)$ in (2.3) with $\Delta \asymp 1$ such that $\pi_0 = \pi_1$, $1/\xi^* \asymp \sigma^2/\lambda$, $1/\xi \asymp (\sigma^2/\lambda)(1 + p/n)$ and $\kappa \asymp 1$. In this case, combining Theorems 7, 9 and 10 reveals that with probability $1 - \mathcal{O}(n^{-1})$,

$$R_x(\widehat{g}_x) - R_z^* \lesssim \left[\frac{K\log n}{n} + \frac{\sigma^2}{\lambda} + \left(\frac{p}{n}\frac{\sigma^2}{\lambda}\right)^2\right]\log n, \quad \text{if } B = \boldsymbol{U}_K; \tag{5.6}$$

$$R_x(\widehat{g}_x) - R_z^* \lesssim \left[\frac{K\log n}{n} + \frac{\sigma^2}{\lambda}\right]\log n, \quad \text{if } B = \widetilde{\boldsymbol{U}}_K. \tag{5.7}$$

We have the following conclusions.

(1) If $p < n$, the two rates above coincide and equal (5.7), whence consistency of both PC-based classifiers requires that $K\log^2 n/n \to 0$ and $\sigma^2\log n/\lambda \to 0$.
(2) If $p > n$, it depends on the signal-to-noise ratio (SNR) λ/σ^2 whether or not consistency of the classifier with $B = \boldsymbol{U}_K$ requires an additional condition.
   (a) If the SNR is large such that

$$\frac{\lambda}{\sigma^2} \gtrsim \min\left\{\left(\frac{p}{n}\right)^2, \frac{p}{\sqrt{nK\log n}}\right\}, \tag{5.8}$$

   the two rates in (5.6) and (5.7) also coincide and equal (5.7). In this case, there is no apparent benefit of using an auxiliary data set.
   (b) For relatively smaller values of SNR that fail (5.8), the effect of using $B = \widetilde{\boldsymbol{U}}_K$ based on an independent data set $\tilde{\boldsymbol{X}}$ is real as evidenced in Figure 1 where we keep λ/σ^2, n and K fixed but let p grow.

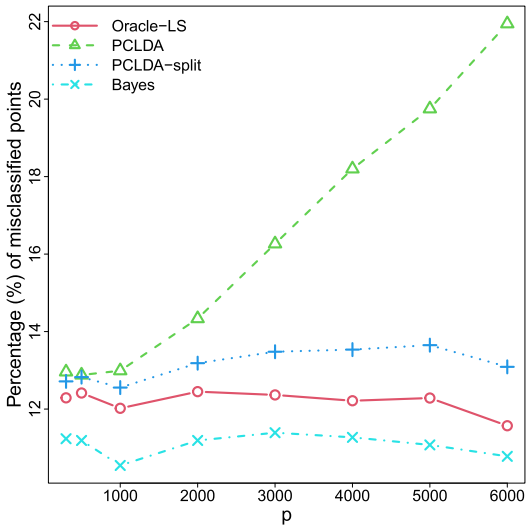

FIG. 1. *Illustration of the advantage of constructing* $\widetilde{\boldsymbol{U}}_K$ *from an independent data set*: *PCLDA represents the PC-based classifier based on* $B = \boldsymbol{U}_K$ *while PCLDA-split uses* $B = \widetilde{\boldsymbol{U}}_K$ *that is constructed from an independent* $\tilde{\boldsymbol{X}}$. *Oracle-LS is the oracle benchmark that uses both* Z *and* $\boldsymbol{Z}$ *while Bayes represents the risk of using the oracle Bayes rule. We fix* $n = 100$ *and* $K = 5$ *and keep* λ/σ^2 *fixed, while we let* p *grow. We refer to Section* 6 *for a detailed data generating mechanism.*

(c) It is worth mentioning that if the SNR is sufficiently large such that

$$\frac{\lambda}{\sigma^2} \gtrsim \max\left\{\left(\frac{p}{n}\right)^2, \frac{p}{\sqrt{nK\log n}}\right\},$$

both errors due to not observing Z and estimation of the column space of the matrix A are negligible compared to the parametric rate K/n, to wit, both rates in (5.6) and (5.7) reduce to $K\log^2 n/n$.

Conditions $\lambda \gtrsim p$ and $\sigma^2 = \mathcal{O}(1)$ are common in the analysis of factor models with a diverging number of features [5, 28, 45]. For instance, $\lambda \gtrsim p$ holds when eigenvalues of $\Sigma_{Z|Y}$ are bounded and a fixed proportion of rows of A are i.i.d. realizations of a sub-Gaussian random vector with covariance matrix having bounded eigenvalues as well. In this case, the bounds in (5.6) and (5.7) reduce to

$$\frac{K\log^2 n}{n} + \frac{\log n}{p},$$

which decreases as p increases. Nevertheless, consistency of the PC-based classifiers only requires $\lambda/\{\sigma^2\log n(1 + p/n)\} \to \infty$ for $B = \boldsymbol{U}_K$ and $\lambda/(\sigma^2\log n) \to \infty$ for $B = \widetilde{\boldsymbol{U}}_K$, which are both much milder conditions.

5.3. *Optimality of the PC-based LDA by sample splitting.* We now show that the PC-based LDA by sample splitting achieves the minimax lower bounds in Theorem 3, up to multiplicative logarithmic factors of n. Recalling that (2.3), for any $\theta \in \Theta(\lambda, \sigma, \Delta)$, one has $\pi_0 = \pi_1$, $1/\xi^* \asymp \sigma^2/\lambda$, $1/\xi \asymp (\sigma^2/\lambda)(1 + p/n)$ and $1 \lesssim \kappa \lesssim 1 + \Delta^2$. Based on Theorem 10, we have the following corollary for the classifier that uses $B = \widetilde{\boldsymbol{U}}_K$. Its proof can be found in Appendix A.4.5. We use the notation $\lessapprox$ for inequalities that hold up to a multiplicative logarithmic factor of n. Recall ω_n^* from (2.5).

COROLLARY 11. *Under model* (1.1), *assume* (i)–(v), $K \log n \le cn$, $\kappa^2\sigma^2/\lambda \le c'$ *and* $\kappa^2\sigma^2 p/(\lambda n) \le c''$ *for some constants* $c, c', c'' > 0$. *For any* $\theta \in \Theta(\lambda, \sigma, \Delta)$, *with probability* $1 - \mathcal{O}(n^{-1})$, *the classifier that uses* $B = \widetilde{\boldsymbol{U}}_K$ *satisfies the following statements*:

(1) *If* $\Delta \asymp 1$, *then*

$$R_x(\widehat{g}_x) - R_z^* \lessapprox (\omega_n^*)^2.$$

(2) *If* $\Delta \to \infty$, *and additionally*, $(\log n + \Delta^2)K \log n/n \to 0$ *and* $(\log n + \Delta^2)\sigma^2/\lambda \to 0$ *as* $n \to \infty$, *then*

$$R_x(\widehat{g}_x) - R_z^* \lessapprox (\omega_n^*)^2 \exp\left\{-\left[\frac{1}{8} + o(1)\right]\Delta^2\right\}.$$

(3) *If* $\Delta \to 0$ *as* $n \to \infty$, *then*

$$R_x(\widehat{g}_x) - R_z^* \lessapprox \min\left\{\frac{\omega_n^*}{\Delta}, 1\right\}\omega_n^*.$$

In view of Theorem 3 and Corollary 11, we conclude the optimality of PC-based procedure that uses $B = \widetilde{\boldsymbol{U}}_K$ over $\Theta(\lambda, \sigma, \Delta)$. For $\Delta \to \infty$, if conditions in (2) are not met such as $\Delta^2 \gtrsim n/K$ or $\Delta^2 \gtrsim \lambda/\sigma^2$, the PC-based procedure still has n^{-a} convergence rate of its excess risk, for arbitrary large $a \ge 1$, as commented in Remark 10.

Regarding the PC-based classifier that does not resort to sample splitting, according to Theorems 3 and 9, its excess risk also achieves optimal rates of convergence when λ/σ^2 is large in the precise sense that

$$\frac{\lambda}{\sigma^2} \gtrsim \min\left\{\frac{1}{\min\{1, \Delta\}}\left(\frac{p}{n}\right)^2, \frac{p}{\sqrt{nK \log n}}\right\}.$$

6. Simulation study. We conduct various simulation studies in this section to compare the performance of our proposed algorithm with other competitors. For our proposed algorithm, we call it PCLDA standing for the Principal Components based LDA. The name PCLDA-K is reserved when the true K is used as input. When K is estimated by $\widehat{K}$, we use PCLDA-$\widehat{K}$ instead. We call PCLDA-CF-k the PCLDA with k-fold cross-fitting. We consider $k = 5$ in our simulation as suggested by [21]. To set a benchmark for PCLDA-CF-k, we use PCLDA-split that uses an independent copy of $\boldsymbol{X}$ to compute $\widetilde{\boldsymbol{U}}_K$. On the other hand, we compare with the nearest shrunken centroids classifier (PAMR) [48], the ℓ_1-penalized linear discriminant (PenalizedLDA) [52] and the direct sparse discriminant analysis (DSDA) [40]. [1] Finally, we choose the performance of the oracle procedure (Oracle-LS) as a benchmark in which Oracle-LS uses both $\boldsymbol{Z}$ and Z to estimate β, β_0 and the classification rule g_z in (3.3).

We generate the data as follows. First, we set $\pi_0 = \pi_1 = 0.5$, $\alpha_0 = \mathbf{0}_K$ and $\alpha_1 = \mathbf{1}_K\sqrt{\eta/K}$. The parameter η controls the signal strength Δ in (2.2). We generate $\Sigma_{Z|Y}$ by independently sampling its diagonal elements $[\Sigma_{Z|Y}]_{ii}$ from Unif(1,3) and set its off-diagonal elements as

$$[\Sigma_{Z|Y}]_{ij} = \sqrt{[\Sigma_{Z|Y}]_{ii}[\Sigma_{Z|Y}]_{jj}}(-1)^{i+j}(0.5)^{|i-j|} \quad \text{for each } i \ne j.$$

The covariance matrix Σ_W is generated in the same way, except we set $\operatorname{diag}(\Sigma_W) = \mathbf{1}_p$. The rows of $\boldsymbol{W} \in \mathbb{R}^{n\times p}$ are generated independently from $N_p(0, \Sigma_W)$. Entries of A are generated independently from $N(0, 0.3^2)$. The training data $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are generated according to models (1.1) and (1.3). In the same way, we generate 100 data points that serve as test data for calculating the (out-of-sample) misclassification error for each algorithm.

[1] PAMR, PenalizedLDA and DSDA are implemented in the R packages pamr, penalizedLDA and TULIP, respectively.

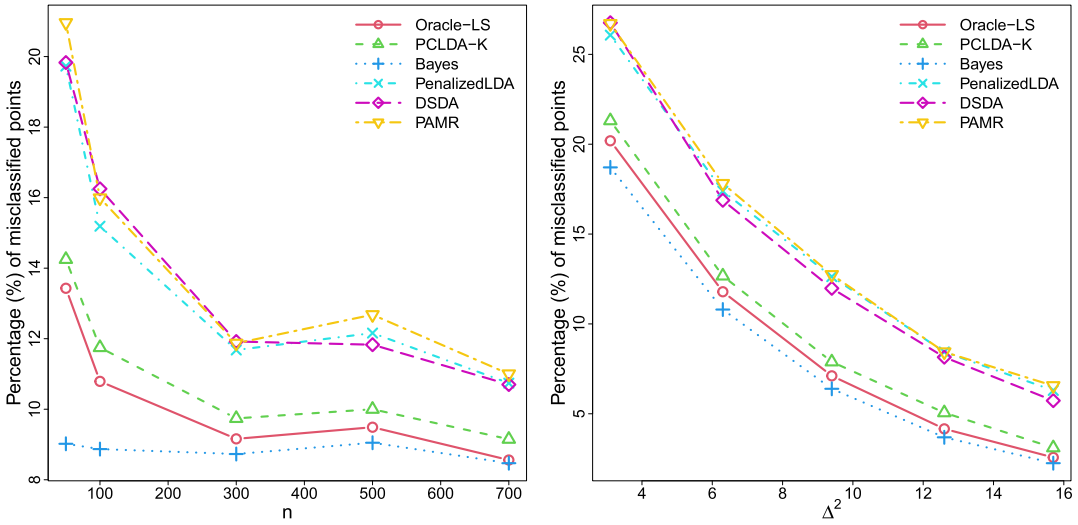

FIG. 2. *The averaged misclassification errors of each algorithm. We vary n in the left panel while we vary Δ in the right panel.*

In the sequel, we vary the dimensions n and p as well as the signal strength Δ in (2.2), one at a time. For each setting, we repeat the entire procedure 100 times and averaged misclassification errors for each algorithm are reported.

6.1. *Vary the sample size n.* We set $\eta = 5$, $K = 10$, $p = 300$ and vary n within $\{50, 100, 300, 500, 700\}$. The left-panel in Figure 2 shows the averaged misclassification error (in percentage) of each algorithm on the test data sets. Since $\widehat{K}$ consistently estimates K, we only report the performance of PCLDA-K. We also exclude the performance of PCLDA-split and PCLDA-CF-5 since they all have similar performance as PCLDA-K.[2] The blue line represents the optimal Bayes error. All algorithms perform better as the sample size n increases. As expected, Oracle-LS is the best because it uses the true $\boldsymbol{Z}$ and Z. Among the other algorithms, PCLDA-K has the closest performance to Oracle-LS in all settings. The gap between PCLDA-K and Oracle-LS does not close as n increases. According to Theorem 9, this is because such a gap mainly depends on $1/\xi$, which does not vary with n.

6.2. *Vary the signal strength Δ^2.* We fix $K = 5$, $n = 100$, $p = 300$ and vary η within $\{2, 4, 6, 8, 10\}$. As a consequence, the signal strength Δ^2 varies within $\{3.1, 6.3, 9.4, 12.6, 15.7\}$. The right panel of Figure 2 depicts the averaged misclassification errors of each algorithm. For the same reasoning as before, we exclude PCLDA-$\widehat{K}$, PCLDA-CF-5 and PCLDA-split. It is evident that all algorithms have better performance as the signal strength Δ increases. Among them, PCLDA-K has the closest performance to Oracle-LS and Bayes in all settings.

6.3. *Vary the feature dimension p.* We examine the performance of each algorithm when the feature dimension p varies across a wide range. Specifically, we fix $K = 5$, $\eta = 5$, $n = 100$ and vary p within $\{100, 300, 500, 700, 900\}$. Figure 3 shows the misclassification errors of each algorithm. The performance of PCLDA-K improves and gets closer to that of Oracle-LS as p increases, in line with Theorem 9. The gap between Oracle-LS and Bayes is due to the fact that both n and Δ are held fixed.

[2]This is as expected since our data generating mechanism ensures $\xi^* \asymp p$ in which case PCLDA-split has no clear advantage comparing to PCLDA-K (see discussions after Theorem 10).

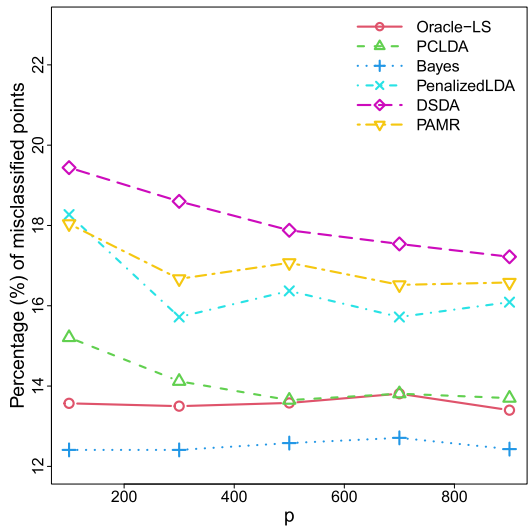

FIG. 3. *The averaged misclassification errors of each algorithm for various choices of* p.

7. Real data analysis. To further illustrate the effectiveness of our proposed method, we analyze three popular gene expression data sets (leukemia data, colon data and lung cancer data),[3] which have been widely used to test classification methods; see, for instance, [2, 25, 42, 44] and also the more recent literature, [19, 27, 40]. These data sets contain thousands or even over ten-thousand features with around one hundred samples (see Table 1). In such challenging settings, LDA-based classifiers that are designed for high-dimensional data not only are easy to interpret but also have competing and even superior performance than other highly complex classifiers such as classifiers based on kernel support vector machines, random forests and boosting [25, 40].

Since the goal is to predict a dichotomous response, for instance, whether one sample is a tumor or normal tissue, we compare the classification performance of each algorithm. For all three data sets, the features are standardized to zero mean and unit standard deviation. For each data set, we randomly split the data, within each category, into 70% training set and 30% test set. Different classifiers are fitted on the training set and their misclassification errors are computed on the test set. This whole procedure is repeated 100 times. The averaged misclassification errors (in percentage) as well as their standard deviations of each algorithm are reported in Table 2. Our proposed PC-based LDA classifiers have the smallest misclassification errors over all data sets. Although PCLDA-CF-5 only has the second best performance in colon and lung cancer data sets, its performance is very close to that of PCLDA-$\widehat{K}$.

TABLE 1
Summary of three data sets

Data name	p	n	n_0 (category)	n_1 (category)
Leukemia	7129	72	47 (acute lymphoblastic leukemia)	25 (acute myeloid leukemia)
Colon	2000	62	22 (normal)	40 (tumor)
Lung cancer	12533	181	150 (adenocarcinoma)	31 (malignant pleural mesothelioma)

[3]Leukemia data is available at www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. Colon data is available from the R package plsgenomics. Lung cancer data is available at www.chestsurg.org.

TABLE 2
The averaged misclassification errors (in percentage). The numbers in parentheses are the standard deviations over 100 repetitions

	PCLDA-$\widehat{K}$	PCLDA-CF-5	DSDA	PenalizedLDA	PAMR
Leukemia	3.57 (0.036)	**3.04** (0.032)	5.52 (0.044)	3.91 (0.043)	4.61 (0.039)
Colon	**16.37** (0.077)	18.11 (0.082)	18.11 (0.07)	33.95 (0.086)	19.00 (0.089)
Lung cancer	**0.55** (0.008)	0.60 (0.009)	1.69 (0.017)	1.80 (0.026)	0.91 (0.011)

8. Extension to multiclass classification. In this section, we discuss how to extend the previously discussed procedure to multiclass classification problems in which Y has L classes, $\mathcal{L} := \{0, 1, \ldots, L-1\}$, for some positive integer $L \geq 2$, and

$$Z|Y = k \sim N_K(\alpha_k, \Sigma_{Z|Y}), \qquad \mathbb{P}(Y = k) = \pi_k, \quad k \in \mathcal{L}. \tag{8.1}$$

In particular, the covariance matrices for the L classes are the same.

For a new point $z \in \mathbb{R}^K$, the oracle Bayes rule assigns it to $k \in \mathcal{L}$ if and only if

$$\begin{aligned} k &= \arg\max_{\ell \in \mathcal{L}} \mathbb{P}(Y = \ell | Z = z) = \arg\max_{\ell \in \mathcal{L}} \log \frac{\mathbb{P}(Z = z, Y = \ell)}{\mathbb{P}(Z = z, Y = 0)} \\ &= \arg\max_{\ell \in \mathcal{L}} \big(z^\top \eta^{(\ell)} + \eta_0^{(\ell)}\big) := \arg\max_{\ell \in \mathcal{L}} G_z^{(\ell|0)}(z), \end{aligned} \tag{8.2}$$

where

$$\eta^{(\ell)} = \Sigma_{Z|Y}^{-1}(\alpha_\ell - \alpha_0), \qquad \eta_0^{(\ell)} = -\frac{1}{2}(\alpha_0 + \alpha_\ell)^\top \eta^{(\ell)} + \log \frac{\pi_\ell}{\pi_0}, \quad \forall \ell \in \mathcal{L}. \tag{8.3}$$

Notice that $G_z^{(0|0)}(z) = 0$ and, for any $\ell \in \mathcal{L} \setminus \{0\}$, the proof of (3.2) reveals that

$$G_z^{(\ell|0)}(z) = z^\top \eta^{(\ell)} + \eta_0^{(\ell)} = \frac{1}{\bar{\pi}_0 \bar{\pi}_\ell [1 - (\alpha_\ell - \alpha_0)^\top \beta^{(\ell)}]} \big(z^\top \beta^{(\ell)} + \beta_0^{(\ell)}\big) \tag{8.4}$$

with $\bar{\pi}_0 = \pi_0/(\pi_0 + \pi_\ell)$, $\bar{\pi}_\ell = \pi_\ell/(\pi_0 + \pi_\ell)$,

$$\begin{aligned} \beta^{(\ell)} &= \big[\mathrm{Cov}\big(Z | Y \in \{0, \ell\}\big)\big]^{-1} \mathrm{Cov}\big(Z, \mathbb{1}\{Y = \ell\} | Y \in \{0, \ell\}\big), \\ \beta_0^{(\ell)} &= -\frac{1}{2}(\alpha_0 + \alpha_\ell)^\top \beta^{(\ell)} + \bar{\pi}_0 \bar{\pi}_\ell \big(1 - (\alpha_\ell - \alpha_0)^\top \beta^{(\ell)}\big) \log \frac{\bar{\pi}_\ell}{\bar{\pi}_0}. \end{aligned} \tag{8.5}$$

In view of (8.2) and (8.4), for a new point $x \in \mathbb{R}^p$ and any matrix $B \in \mathbb{R}^{p \times q}$ with $q \in [p]$, we propose the following multiclass classifier

$$\widehat{g}_x^*(x) = \arg\max_{\ell \in \mathcal{L}} \widehat{G}_x^{(\ell|0)}(x), \tag{8.6}$$

where $\widehat{G}_x^{(0|0)}(x) = 0$ and, for any $\ell \in \mathcal{L} \setminus \{0\}$,

$$\widehat{G}_x^{(\ell|0)}(x) = \frac{1}{\widetilde{\pi}_0 \widetilde{\pi}_\ell [1 - (\widehat{\mu}_\ell - \widehat{\mu}_0)^\top \widehat{\theta}^{(\ell)}]} \big(x^\top \widehat{\theta}^{(\ell)} + \widehat{\beta}_0^{(\ell)}\big) \tag{8.7}$$

with

$$\begin{aligned} \widetilde{\pi}_\ell &= \frac{n_\ell}{n_0 + n_\ell}, \\ \widehat{\theta}^{(\ell)} &= B\big(\Pi_{(n_0+n_\ell)} \boldsymbol{X}^{(\ell)} B\big)^+ \boldsymbol{Y}^{(\ell)}, \\ \widehat{\beta}_0^{(\ell)} &= -\frac{1}{2}(\widehat{\mu}_0 + \widehat{\mu}_\ell)^\top \widehat{\theta}^{(\ell)} + \widetilde{\pi}_0 \widetilde{\pi}_\ell \big(1 - (\widehat{\mu}_\ell - \widehat{\mu}_0)^\top \widehat{\theta}^{(\ell)}\big) \log \frac{\widetilde{\pi}_\ell}{\widetilde{\pi}_0}. \end{aligned}$$

Here, n_ℓ and $\widehat{\mu}_\ell$ are the nonparametric estimates as (3.6) and both the submatrix $\boldsymbol{X}^{(\ell)} \in \mathbb{R}^{(n_0+n_\ell)\times p}$ of $\boldsymbol{X}$ and the response vector $\boldsymbol{Y}^{(\ell)} = \{0,1\}^{(n_0+n_\ell)}$ correspond to samples with label in $\{0,\ell\}$. Note that $\boldsymbol{Y}^{(\ell)}$ is encoded as 1 for observations with label ℓ and 0 otherwise.

To analyze the classifier $\widehat{g}_x^*$ in (8.6), its excess risk depends on

$$\widehat{r}_1 = \max_{\ell\in\mathcal{L}\setminus\{0\}} \big\|[\Sigma_Z^{(\ell)}]^{1/2}(A^\top\widehat{\theta}^{(\ell)} - \beta^{(\ell)})\big\|_2, \qquad \widehat{r}_2 = \max_{\ell\in\mathcal{L}\setminus\{0\}} \big\|\widehat{\theta}^{(\ell)}\big\|_2 \tag{8.8}$$

as well as $\widehat{r}_3$ as defined in (4.7). Here, $\Sigma_Z^{(\ell)} := \mathrm{Cov}(Z|Y \in \{0,\ell\})$. Analogous to (4.9), for some constant $C = C(\gamma) > 0$, define

$$\widehat{\omega}_n = C\sqrt{\log n}\bigg(\widehat{r}_1 + \|\Sigma_W\|_{\mathrm{op}}^{1/2}\widehat{r}_2 + \widehat{r}_2\widehat{r}_3 + \sqrt{\frac{L}{n}}\bigg). \tag{8.9}$$

For ease of presentation, we also assume there exists some sequence $\Delta > 0$ and some absolute constants $C > c > 0$ such that

$$c\Delta \le \min_{k,\ell\in\mathcal{L},k\neq\ell} \|\alpha_\ell - \alpha_k\|_{\Sigma_{Z|Y}} \le \max_{k,\ell\in\mathcal{L},k\neq\ell} \|\alpha_\ell - \alpha_k\|_{\Sigma_{Z|Y}} \le C\Delta. \tag{8.10}$$

The following theorem extends Theorem 7 to multiclass classification by establishing rates of convergence of the excess risk of $\widehat{g}_x^*$ in (8.6) for a general $B \in \mathbb{R}^{p\times q}$.

THEOREM 12. *Under model* (1.1) *and* (8.1), *assume* (i)–(iii) *and* (8.10). *Further, assume* $c/L \le \min_{k\in\mathcal{L}}\pi_k \le \max_{k\in\mathcal{L}}\pi_k \le C/L$ *and* $LK\log n \le c'n$ *for some constants* $c, c', C > 0$. *Then, for any sequence* $\omega_n > 0$ *satisfying* $(1+\Delta^2)\omega_n = o(1)$ *as* $n\to\infty$, *on the event* $\{\widehat{\omega}_n \le \omega_n\}$, *the following holds with probability at least* $1 - \mathcal{O}(n^{-1})$ *under the law* $\mathbb{P}^{\boldsymbol{D}}$:

(1) *If* $\Delta \asymp 1$, *then*

$$R_x(\widehat{g}_x^*) - R_z^* \lesssim L\omega_n^2.$$

(2) *If* $\Delta\to\infty$, *then, for some constant* $c'' > 0$,

$$R_x(\widehat{g}_x^*) - R_z^* \lesssim L\omega_n^2\exp\{-[c'' + o(1)]\Delta^2\}$$

(3) *If* $\Delta = o(1)$, *then*

$$R_x(\widehat{g}_x^*) - R_z^* \lesssim L\omega_n\min\bigg\{\frac{\omega_n}{\Delta}, 1\bigg\}.$$

PROOF. The proof can be found in Appendix A.5. □

Condition (8.10) is only assumed to simplify the presentation. It is straightforward to derive results based on our analysis when the separation $\|\alpha_\ell - \alpha_k\|_{\Sigma_{Z|Y}}$ is not of the same order for all $\ell, k \in \mathcal{L}$. For the third case, $\Delta = o(1)$, our proof also allows to establish different convergence rates depending on whether or not π_k and π_ℓ are distinct for each $k \neq \ell$, analogous to the last two cases of Theorem 7. However, we opt for the current presentation for succinctness.

Theorem 12 immediately leads to the following corollary for the PC-based classifiers that use $B = \boldsymbol{U}_K$ and $B = \widetilde{\boldsymbol{U}}_K$. Furthermore, Theorem 8 also ensures that similar guarantees can be obtained for the classifiers in (8.6) that use $B = \boldsymbol{U}_{\widehat{K}}$ and $B = \widetilde{\boldsymbol{U}}_{\widetilde{K}}$.

COROLLARY 13. *Assume the conditions in Theorem* 12 *and* $\xi \ge C\kappa^2$ *for some constant* $C > 0$. *Then the conclusion of Theorem* 12 *holds for the classifier in* (8.6) *that uses*:

(1) $B = \boldsymbol{U}_K$ *with*

$$\omega_n = \Big(\sqrt{\frac{LK\log n}{n}} + \min\{1, \Delta\}\sqrt{\frac{1}{\xi^*}} + \sqrt{\frac{\kappa}{\xi^2}}\Big)\sqrt{\log n},$$

(2) $B = \widetilde{\boldsymbol{U}}_K$ *with*

$$\omega_n = \Big(\sqrt{\frac{LK\log n}{n}} + \min\{1, \Delta\}\sqrt{\frac{1}{\xi^*}}\Big)\sqrt{\log n}.$$

PROOF. See Appendix A.5.3. □

REMARK 13. Multiclass classification problems based on discriminant analysis have been studied, for instance, by [20, 23, 39, 52]. Theoretical guarantees are only provided in [39] and [20] under the classical LDA setting for moderate/large separation scenarios, $\Delta \gtrsim 1$, and for fixed L, the number of classes; see also the work [1] that derives bounds for the misclassification error (rather than excess risk) in a set-up similar to LDA, and reports a similar phase transition phenomenon between $\Delta \asymp 1$ and $\Delta \to \infty$. Our results fully characterize dependence of the excess risk on L and also cover the weak separation case, $\Delta \to 0$.

REMARK 14. The classifier in (8.6) chooses $Y = 0$ as the baseline. In practice, we recommend taking each class as the baseline one at the time and averaging the predicted probabilities. Specifically, it is easy to see that, for any baseline choice $k \in \mathcal{L}$ and for any $\ell \in \mathcal{L}$,

$$\mathbb{P}(Y = \ell | Z = z) = \frac{\mathbb{P}(Z = z, Y = \ell)}{\sum_{k' \in \mathcal{L}} \mathbb{P}(Z = z, Y = k')} = \frac{\exp\{G_z^{(\ell|k)}(z)\}}{\sum_{k' \in \mathcal{L}} \exp\{G_z^{(k'|k)}(z)\}},$$

where $G_z^{(\ell|k)}(z)$ is defined analogous to (8.2) with k in lieu of 0. Therefore, for any new data point $x \in \mathbb{R}^p$, the averaged version of the classifier in (8.6) is

$$\arg\max_{\ell \in \mathcal{L}} \frac{1}{L} \sum_{k \in \mathcal{L}} \frac{\exp\{\widehat{G}_x^{(\ell|k)}(x)\}}{\sum_{k' \in \mathcal{L}} \exp\{\widehat{G}_x^{(k'|k)}(x)\}}$$

with $\widehat{G}_x^{(\ell|k)}(x)$ defined analogous to (8.7). This classifier tends to have better finite sample performance, as revealed by the simulation study in Appendix E.3.

Acknowledgments. The authors would like to thank the Editor, Associate Editor and two referees for their careful reading and very constructive suggestions.

Funding. Wegkamp is supported in part by the National Science Foundation grants DMS 2015195 and DMS 2210557. Bing is partially supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal discriminant analysis in high-dimensional latent factor models" (DOI: 10.1214/23-AOS2289SUPP; .pdf). Appendices A and B contain the main proofs for the results in Sections 2–5 and 8. Technical lemmas and auxiliary lemmas are collected in Appendices C and D. Appendix E contains additional simulation results.

REFERENCES

[1] ABRAMOVICH, F. and PENSKY, M. (2019). Classification with many classes: Challenges and pluses. *J. Multivariate Anal.* **174** 104536. MR3995262 https://doi.org/10.1016/j.jmva.2019.104536

[2] 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. *Proc. Natl. Acad. Sci. USA* **96** 6745–6750.

[3] ANTONIADIS, A., LAMBERT-LACROIX, S. and LEBLANC, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* **19** 563–570.

[4] AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) **26**. Curran Associates, Red Hook.

[5] BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313 https://doi.org/10.1214/11-AOS966

[6] BAI, J. and NG, S. (2008). Forecasting economic time series using targeted predictors. *J. Econometrics* **146** 304–317. MR2465175 https://doi.org/10.1016/j.jeconom.2008.08.010

[7] BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436 https://doi.org/10.1198/016214505000000628

[8] BARKER, M. and RAYENS, W. (2003). Partial least squares for discrimination. *J. Chemom.* **17** 166–173.

[9] BARSHAN, E., GHODSI, A., AZIMIFAR, Z. and JAHROMI, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit.* **44** 1357–1371.

[10] BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131 https://doi.org/10.3982/ECTA9626

[11] BIAU, G., BUNEA, F. and WEGKAMP, M. H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory* **51** 2163–2172. MR2235289 https://doi.org/10.1109/TIT.2005.847705

[12] BING, X., BUNEA, F., NING, Y. and WEGKAMP, M. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering. *Ann. Statist.* **48** 2055–2081. MR4134786 https://doi.org/10.1214/19-AOS1877

[13] BING, X., BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2021). Prediction under latent factor regression: Adaptive PCR, interpolating predictors and beyond. *J. Mach. Learn. Res.* **22** Paper No. 177. MR4318533 https://doi.org/10.22405/2226-8383-2021-22-1-177-187

[14] BING, X. and WEGKAMP, M. (2022). Interpolating discriminant functions in high-dimensional Gaussian latent mixtures. Available at arXiv:2210.14347.

[15] BING, X. and WEGKAMP, M. (2023). Supplement to "Optimal discriminant analysis in high-dimensional latent factor models." https://doi.org/10.1214/23-AOS2289SUPP

[16] BING, X. and WEGKAMP, M. H. (2019). Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.* **47** 3157–3184. MR4025738 https://doi.org/10.1214/18-AOS1774

[17] BOULESTEIX, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 33. MR2101480 https://doi.org/10.2202/1544-6115.1075

[18] CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857 https://doi.org/10.1198/jasa.2011.tm11199

[19] CAI, T. T. and ZHANG, L. (2019). High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 675–705. MR3997097

[20] CAI, T. T. and ZHANG, L. (2021). A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *Ann. Statist.* **49** 1537–1568. MR4298872 https://doi.org/10.1214/20-aos2012

[21] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[22] CHIAROMONTE, F. and MARTINELLI, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.* **176** 123–144. MR1869195 https://doi.org/10.1016/S0025-5564(01)00106-7

[23] CLEMMENSEN, L., HASTIE, T., WITTEN, D. and ERSBØLL, B. (2011). Sparse discriminant analysis. *Technometrics* **53** 406–413. MR2850472 https://doi.org/10.1198/TECH.2011.08118

[24] DAI, J. J., LIEU, L. and ROCKE, D. (2006). Dimension reduction for classification with gene expression microarray data. *Stat. Appl. Genet. Mol. Biol.* **5** Art. 6. MR2221299 https://doi.org/10.2202/1544-6115.1147

[25] DETTLING, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583–3593. https://doi.org/10.1093/bioinformatics/bth447

[26] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[27] FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009 https://doi.org/10.1214/07-AOS504

[28] FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016

[29] FAN, J., XUE, L. and YAO, J. (2017). Sufficient forecasting using factor models. *J. Econometrics* **201** 292–306. MR3717565 https://doi.org/10.1016/j.jeconom.2017.08.009

[30] GHOSH, D. (2001). Singular value decomposition regression models for classification of tumors from microarray experiments. In *Biocomputing* 2002 18–29. World Scientific, Singapore.

[31] HADEF, H. and DJEBABRA, M. (2019). Proposal method for the classification of industrial accident scenarios based on the improved principal components analysis (improved PCA). *Prod. Eng.* **13** 53–60.

[32] HAHN, P. R., CARVALHO, C. M. and MUKHERJEE, S. (2013). Partial factor modeling: Predictor-dependent shrinkage for linear regression. *J. Amer. Statist. Assoc.* **108** 999–1008. MR3174679 https://doi.org/10.1080/01621459.2013.779843

[33] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

[34] HOTELLING, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *Br. J. Stat. Psychol.* **10** 69–79.

[35] IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques*: *Regression*, *Classification*, *and Manifold Learning*. *Springer Texts in Statistics*. Springer, New York. MR2445017 https://doi.org/10.1007/978-0-387-78189-1

[36] JIN, D., HENRY, P., SHAN, J. and CHEN, J. (2021). Classification of cannabis strains in the Canadian market with discriminant analysis of principal components using genome-wide single nucleotide polymorphisms. *PLoS ONE* **16** e0253387.

[37] LI, H. (2016). Accurate and efficient classification based on common principal components analysis for multivariate time series. *Neurocomputing* **171** 744–753.

[38] MA, Z., LIU, Z., ZHAO, Y., ZHANG, L., LIU, D., REN, T., ZHANG, X. and LI, S. (2020). An unsupervised crop classification method based on principal components isometric binning. *ISPRS Int.l J. Geo-Inf.* **9** 648.

[39] MAI, Q., YANG, Y. and ZOU, H. (2019). Multiclass sparse discriminant analysis. *Statist. Sinica* **29** 97–111. MR3889359

[40] MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661 https://doi.org/10.1093/biomet/asr066

[41] MALLARY, C., BERG, C., BUCK, J. R., TANDON, A. and ANDONIAN, A. (2022). Acoustic rainfall detection with linear discriminant functions of principal components. *J. Acoust. Soc. Am.* **151** A149–A149.

[42] NGUYEN, D. V. and ROCKE, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18** 39–50.

[43] SHAO, J., WANG, Y., DENG, X. and WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39** 1241–1265. MR2816353 https://doi.org/10.1214/10-AOS870

[44] SINGH, D., FEBBO, P. G., ROSS, K., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D'AMICO, A. V. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1** 203–209. https://doi.org/10.1016/s1535-6108(02)00030-2

[45] STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271 https://doi.org/10.1198/016214502388618960

[46] STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257 https://doi.org/10.1198/073500102317351921

[47] TARIGAN, B. and VAN DE GEER, S. A. (2006). Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli* **12** 1045–1076. MR2274857 https://doi.org/10.3150/bj/1165269150

[48] TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99** 6567–6572.

[49] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

[50] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452 https://doi.org/10.1214/13-AOS1151

[51] WEGKAMP, M. and YUAN, M. (2011). Support vector machines with a reject option. *Bernoulli* **17** 1368–1385. MR2854776 https://doi.org/10.3150/10-BEJ320

[52] WITTEN, D. M. and TIBSHIRANI, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 753–772. MR2867457 https://doi.org/10.1111/j.1467-9868.2011.00783.x