Journal of Multivariate Analysis 167 (2018) 269–283

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A U-classifier for high-dimensional data under non-normality

M. Rauf Ahmad *, Tatjana Pavlenko

Department of Statistics, Uppsala University, Sweden
Department of Mathematics, KTH, Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 2 October 2017
Available online 28 May 2018

AMS subject classification:
62H30

Keywords:
Bias-adjusted classifier
High-dimensional classification
U-statistics

ABSTRACT

A classifier for two or more samples is proposed when the data are high-dimensional and the distributions may be non-normal. The classifier is constructed as a linear combination of two easily computable and interpretable components, the U-component and the P-component. The U-component is a linear combination of U-statistics of bilinear forms of pairwise distinct vectors from independent samples. The P-component, the discriminant score, is a function of the projection of the U-component on the observation to be classified. Together, the two components constitute an inherently bias-adjusted classifier valid for high-dimensional data. The classifier is linear but its linearity does not rest on the assumption of homoscedasticity. Properties of the classifier and its normal limit are given under mild conditions. Misclassification errors and asymptotic properties of their empirical counterparts are discussed. Simulation results are used to show the accuracy of the proposed classifier for small or moderate sample sizes and large dimensions. Applications involving real data sets are also included.

1. Introduction

A linear classifier for $g \geq 2$ populations is presented when the data are high-dimensional and possibly non-normal. For each $i \in \{1, \ldots, g\}$, let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ be n_i independent and identically distributed random vectors from the ith population with distribution function $\mathcal{F}_i$. It is assumed that, for all $k \in \{1, \ldots, n_i\}$, $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})^\top$ with mean vector $\mathrm{E}(\mathbf{x}_{ik}) = \boldsymbol{\mu}_i$ and covariance matrix $\mathrm{cov}(\mathbf{x}_{ik}) = \boldsymbol{\Sigma}_i > 0$. We are interested in constructing a classifier for high-dimensional, low sample size setting, i.e., $p \gg n_i$, where $\mathcal{F}_i$ need not be normal.

Classification and regression are two of the most powerful tools of statistical analysis, both as the main objective of analysis on their own and as a source of further investigation. Due to the ever growing complexity of data, classification has attracted a central place in modern statistical analysis. The wave of large-dimensional data in the last few decades and associated questions have led researchers to improve substantially the classical theory of classification.

This paper mainly addresses the classification problem for such a complex data set up, particularly when the dimension p of the multivariate vector exceeds the number n_i of such vectors, i.e., $p \gg n_i$; see Section 5. As the classical theory of classification does not work in this case, mainly due to the singularity of the empirical covariance matrix (see Section 2 for more details), efforts have been made in the literature to offer potential alternatives. Bickel and Levina [9] discuss the Independence Rule (IR), or naive Bayes rule, by using only the diagonal of the empirical covariance matrix and compare it to Fisher's linear discriminant function (LDF) for the case of two normal populations. Under certain conditions on the eigenvalues of the scaled covariance matrix, they show that IR, under the assumption of independence, is comparable to Fisher's LDF under dependence when the empirical covariance matrix is replaced with a g-inverse computed from the

* Correspondence to: Department of Statistics, Uppsala University, Ekonomikum, Kyrkogårdsgatan 10, Box 513, 75120, Uppsala, Sweden.
E-mail address: rauf.ahmad@statistik.uu.se (M. Rauf Ahmad).

https://doi.org/10.1016/j.jmva.2018.05.008
0047-259X/

empirical non-zero eigenvalues and the corresponding eigenvectors. See also [15,19] for discriminant analysis based on a diagonal covariance matrix, and [38] where the regular inverse of a regularized covariance estimator is used.

Most of the modifications of the classical theory pertain to the linear classifier, with particular focus on sparsity. For some recent attempts, see [10,12,18,20,27,34]. A regularized discriminant analysis using Fisher's LDF is given in [39]. Classifier performance measures for high-dimensional data are discussed in [17]. Recently, there have also been attempts to extend the quadratic classifier to the high-dimensional case, particularly under sparsity; see, e.g., [16,21,26]. For a review of classification methods for high-dimensional data, see [28].

We begin, in Section 2, with the two-sample U-classifier, giving details on its construction and justification. A multi-sample extension is given in Section 3. Section 4 reports the results of a simulation study and Section 5 describes real data applications. Technical proofs are deferred to the Appendix.

2. The two-sample case

2.1. Construction and motivation of the U-classifier

For each $i \in \{1, 2\}$, let $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})^\top \sim \mathcal{F}_i$ be as defined above and π_i denote the ith (unknown) population. Let also $\mathcal{R}_i = \{\mathbf{x} : \mathbf{x} \in \pi_i\}$ be the region of observed data from the ith population, where $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathcal{X}$, $\mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset$ with $\mathcal{X}$ the space of observed $\mathbf{x}$, $\emptyset$ the empty set. Further let $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ be the set of parameters for $\mathcal{F}_i$. We denote by $\pi(1|2)$ the error of misclassifying $\mathbf{x}$ in π_1 when it actually comes from π_2. Formally,

$$\pi(1|2) = \Pr(\mathbf{x} \in \mathcal{R}_1 | \mathbf{x} \in \pi_2) = \int_{\mathcal{R}_1} d\mathcal{F}_2(\mathbf{x}|\boldsymbol{\theta}), \tag{1}$$

where $\pi(2|1) = 1 - \pi(1|2)$ is the opposite error of misclassification.

Using the information on p characteristics, $(x_{ik1}, \ldots, x_{ikp})^\top$, our aim is to construct a classifier which assigns $\mathbf{x}$ to π_i optimally, i.e., by keeping $\pi(i|j)$ as small as possible for all $i, j \in \{1, 2\}$, $i \neq j$. As this aim is not achievable, since $\mathcal{F}_i$ or related parameters are unknown, we focus on minimizing an appropriately defined empirical measure for the proposed classifier as $n_i, p \to \infty$ where p may arbitrarily exceed n_i ($p \gg n_i$), $\mathcal{F}_i$ may not necessarily be normal, and $\boldsymbol{\Sigma}_i$ may be unequal. We keep the high-dimensional or (n_i, p)-asymptotic framework general, in that we let $n_i \to \infty$ and $p \to \infty$ but without requiring the two indices to satisfy any relationship of mutual growth order.

To proceed with the idea, let

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{x}_{ik} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1}\sum_{k=1}^{n_i} (\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)^\top \tag{2}$$

be the unbiased estimators of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, respectively. The classical two-sample linear classifier, assuming equal and known $\boldsymbol{\Sigma}$ with equal misclassification costs and population priors, is expressed, ignoring the constants, as

$$C(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - (\bar{\mathbf{x}}_1^\top \boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_2)/2,$$

where $\mathbf{x}$ denotes the point to be classified; see Chapter 6 in [32].

The rule is: $\mathbf{x} \in \pi_1$ if $C(\mathbf{x}) > 0$, else $\mathbf{x} \in \pi_2$. Although $C(\mathbf{x})$ is usually constructed under a normality assumption, using the ratio of multivariate normal density functions and assuming $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Fisher constructed the same classifier without assuming normality, and hence it is also known as Fisher's linear discriminant function. It is the most frequently used classifier and, assuming $n_i > p$ and normality, its misclassification probability can be computed using the normal distribution. Obviously, with $\boldsymbol{\Sigma}$ unknown in practice, we need to estimate $C(\mathbf{x})$, replacing $\boldsymbol{\Sigma}$ with the pooled estimate

$$\widehat{\boldsymbol{\Sigma}}_{\text{pooled}} = \sum_{i=1}^{2}(n_i - 1)\widehat{\boldsymbol{\Sigma}}_i \Big/ \sum_{i=1}^{2}(n_i - 1),$$

where $\widehat{\boldsymbol{\Sigma}}_i$ are defined above, so that

$$\widehat{C}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \widehat{\boldsymbol{\Sigma}}_{\text{pooled}}^{-1}\mathbf{x} - (\bar{\mathbf{x}}_1^\top \widehat{\boldsymbol{\Sigma}}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \widehat{\boldsymbol{\Sigma}}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_2)/2. \tag{3}$$

When the data are high-dimensional, i.e., $p > n_i$, $\widehat{\boldsymbol{\Sigma}}_i$, and hence, $\widehat{\boldsymbol{\Sigma}}_{\text{pooled}}$, is singular so that $\widehat{C}(\mathbf{x})$ cannot be used. To see how the situation develops in this framework, let us first take $\widehat{\boldsymbol{\Sigma}}$ out of the classifier in (3) and consider

$$\widetilde{C}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{x} - (\bar{\mathbf{x}}_1^\top \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \bar{\mathbf{x}}_2)/2. \tag{4}$$

From Chapter 2 in [31], note that $\mathrm{E}(\bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_i) = B_i + \|\boldsymbol{\mu}_i\|^2$, where $\|\mathbf{a}\|^2 = \mathbf{a}^\top \mathbf{a}$ is the Euclidean norm of vector $\mathbf{a}$ and $B_i = \mathrm{tr}(\boldsymbol{\Sigma}_i)/n_i$ for all $i \in \{1, 2\}$. Assuming $\mathbf{x} \in \pi_1$, we have

$$\mathrm{E}\{\widetilde{C}(\mathbf{x})|\mathbf{x} \in \pi_1\} = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/2 - B, \tag{5}$$

with $B = (B_1 - B_2)/2$. We observe that, depriving the classifier of the covariance matrix makes it biased with bias term B composed of the traces of unknown covariance matrices. If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, then $B = (n_2 - n_1)\,\mathrm{tr}(\boldsymbol{\Sigma})/(2n_1n_2)$, so that the classifier is positively (negatively) biased given $n_2 > n_1$ ($n_2 < n_1$), and unbiased if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and $n_1 = n_2$. To inherently adjust it for bias and improve its accuracy, consider the second component of $\widetilde{C}(\mathbf{x})$ in (4) with

$$\mathrm{E}\{(\bar{\mathbf{x}}_1^\top\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top\bar{\mathbf{x}}_2)/2\} = B + (\|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2)/2.$$

Now, write

$$\bar{\mathbf{x}}_i^\top\bar{\mathbf{x}}_i = \frac{1}{n_i^2}\sum_{k=1}^{n_i}\sum_{r=1}^{n_i}A_{ikr} = \frac{1}{n_i^2}\sum_{k=1}^{n_i}A_{ik} + \frac{1}{n_i^2}\sum_{\substack{k=1\\k\neq r}}^{n_i}\sum_{r=1}^{n_i}A_{ikr} = Q_{0i} + Q_{1i},$$

where $A_{ik} = \mathbf{x}_{ik}^\top\mathbf{x}_{ik}$, $A_{ikr} = \mathbf{x}_{ik}^\top\mathbf{x}_{ir}$, $k \neq r$, where $\mathrm{E}(Q_{0i}) = B_i + R_i$ and $\mathrm{E}(Q_{1i}) = \|\boldsymbol{\mu}_i\|^2 - R_i$ with $R_i = \|\boldsymbol{\mu}_1\|^2/n_i$. Denoting $Q_0 = Q_{01} - Q_{02}$, $Q_1 = Q_{11} - Q_{12}$, $R = R_1 - R_2$, we have

$$\mathrm{E}(Q_0) = 2B + R, \quad \mathrm{E}(Q_1) = (\|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2) - R.$$

Adjusting both components for R gives $\mathrm{E}(Q_0) - R = 2B$, $\mathrm{E}(Q_1) + R = \|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2$. This leads to an unbiased version of $\widetilde{C}(\mathbf{x})$ in (4), to be denoted heretofore $A(\mathbf{x})$, as

$$A(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top\mathbf{x}/p - (U_{n_1} - U_{n_2})/2, \tag{6}$$

where $U_{n_i} = \sum_{k\neq r}^{n_i}A_{ikr}/pQ(n_i)$ is a one-sample U-statistic with symmetric kernel [33], $Q(n_i) = n_i(n_i - 1)$ and $A_{ikr}/p = \mathbf{x}_{ik}^\top\mathbf{x}_{ir}/p$, $k \neq r$, is a bilinear form of independent components.

Assuming $\mathbf{x} \in \pi_1$ and independent of both samples, $\mathrm{E}\left\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top\mathbf{x}\right\} = \|\boldsymbol{\mu}_1\|^2 - \boldsymbol{\mu}_2^T\boldsymbol{\mu}_1$ Thus

$$\mathrm{E}\{A(\mathbf{x})|\pi_1\} = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/(2p),$$

without bias term. Further, with $\mathbf{x} \in \pi_1$, $A(\mathbf{x})$ is composed of bilinear forms, two from sample 1, one from sample 2, one mixed. By symmetry, $\mathrm{E}\{A(\mathbf{x})|\pi_2\} = -\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/2p$, with $A(\mathbf{x})$ composed of two bilinear forms from sample 2, one from sample 1, one mixed. We, therefore, define the classification rule for the proposed U-classifier in (6) as

$$\text{Assign } \mathbf{x} \text{ to } \pi_1 \text{ if } A(\mathbf{x}) > 0, \quad \text{otherwise to } \pi_2.$$

Before we study the properties of $A(\mathbf{x})$ in Section 2.2, a few important remarks are in order. First, $A(\mathbf{x})$ is composed of bilinear forms — and we call it *bilinear classifier* — where the bi-linearity of the U-component is expressed in the kernels of the two U-statistics and that of the P-component by the projection of the new observation with respect to the difference between the empirical centroids of the two independent samples. Further, $A(\mathbf{x})$ is entirely composed of empirical quantities, free of any unknown parameter, so that it can be directly used in practice. Moreover, $A(\mathbf{x})$ is linear but the linearity does not rely on a homoscedasticity assumption: it is linear even if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. In the classical case, the violation of homoscedasticity makes the classifier quadratic, which is not the case for $A(\mathbf{x})$. This provides an additional advantage for the proposed classifier so that it can be used without assuming or testing homoscedasticity. The same advantage will be valid for the multi-sample extension in Section 3 as well.

In this context, it may also be mentioned that the theory of U-statistics has recently gained momentum in its use in high-dimensional inference. Although most applications have so far been restricted to testing hypotheses for large dimensional parameters (see, e.g., [1,2] and the references therein), it has recently also been used in classification and cluster analysis. For a general overview and application in genetics, see [13].

Moreover, the first part of $A(\mathbf{x})$ is normalized by p, and so are the kernels of the U-statistics in the second part. This will help us derive the limit distribution of $A(\mathbf{x})$ for (n_i, p)-asymptotics under a general multivariate model and mild assumptions. As a final remark, recall that the formulation of $A(\mathbf{x})$ arises from depriving the original classifier of its empirical covariance matrix. Although removing an essential ingredient has its price, the resulting classifier can still be justified.

First, a completely affine-invariant classifier (or a test statistic) in high-dimensional case is not possible. Given this, scale invariance is mostly compromised due to the singularity of the empirical covariance matrix, as explained above. Thus, a location-invariant classifier is the best that can be achieved. One can use a regularized estimator or a diagonal matrix to keep scale-invariance but such alternatives obviously have their own price, e.g., in terms of loss of huge amount of information. We take the first route and sacrifice scale-invariance for the sake of location-invariance. Second, the proposed classifier can be justified simply as a different type of classifier on its own.

To see this, let $\bar{d}_{12} = \bar{d}_1 - \bar{d}_2$, $\bar{d}_i = (d_{k1} + \cdots + d_{kn_i})/n_i$, where $d_{ki} = \|\mathbf{x} - \mathbf{x}_{ki}\|^2$ is the Euclidean distance of $\mathbf{x}$ from sample $i \in \{1, 2\}$, $k \in \{1, \ldots, n_i\}$. It follows that $\bar{d}_{12}$ has the same bias B as for $A(\mathbf{x})$. Now consider

$$D(\mathbf{x}) = \bar{d}_{12} - \{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_1)/n_1 - \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_2)/n_2\},$$

where

$$\widehat{\boldsymbol{\Sigma}}_i = \sum_{k\neq r}^{n_i}\mathbf{D}_{ikr}^\top\mathbf{D}_{ikr}/n_i(n_i - 1) \quad \text{with } \mathbf{D}_{ikr} = \mathbf{x}_{ik} - \mathbf{x}_{ir};$$

see [3]. As $d_{k1} + \cdots + d_{kn_i} = (n_i - 1)\,\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_i) + n_i\|\mathbf{x} - \bar{\mathbf{x}}_i\|^2$, it simplifies to $D(\mathbf{x}) = A(\mathbf{x}) + B$; see Eq. (5). This implies that $A(\mathbf{x})$ can also be constructed as a distance-based classifier with same bias-adjustment that leads Eqs. (5)–(6). This approach is discussed in [5,11]; see also [6]. Our approach, however, makes the classifier not only unbiased but also more general and practically convenient.

2.2. Asymptotic distribution of the U-classifier

For $i \in \{1, 2\}$ and given $\mathbf{x}_{ik} \sim \mathcal{F}_i$, let $\mathbf{z}_{ik} = \mathbf{x}_{ik} - \boldsymbol{\mu}_i$ with $\mathrm{E}(\mathbf{z}_{ik}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{z}_{ik}) = \boldsymbol{\Sigma}_i$. When we relax normality, we assume the general multivariate model

$$\mathbf{z}_{ik} = \boldsymbol{\Lambda}_i \mathbf{y}_{ik}, \tag{7}$$

where $\mathbf{y}_{ik} = (y_{ik1}, \ldots, y_{ikp})^\top$ has i.i.d elements with $\mathrm{E}(\mathbf{y}_{ik}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{y}_{ik}) = \mathbf{I}$, and $\boldsymbol{\Lambda}_i$ is a known $p \times p$ matrix of constants with $\boldsymbol{\Lambda}_i^\top \boldsymbol{\Lambda}_i = \mathbf{A}_i$, $\boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^\top = \boldsymbol{\Sigma}_i > 0$. For the properties of $A(\mathbf{x})$, we need following assumptions under Model (7).

Assumption 1. $\mathrm{E}(y_{iks}^4) = \gamma < \infty$, $\gamma \in \mathbb{R}^+$ for all $i \in \{1, 2\}$.

Assumption 2. $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i)/p = O(1)$ for all $i \in \{1, 2\}$.

Assumption 3. $\lim_{p\to\infty} \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_k \boldsymbol{\mu}_j/p = O(1)$ for all $i, j, k \in \{1, 2\}$.

Assumption 4. $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i^a \odot \boldsymbol{\Sigma}_j^b)/\mathrm{tr}(\boldsymbol{\Sigma}_i^a \otimes \boldsymbol{\Sigma}_j^b) = 0$ for all $i, j \in \{1, 2\}$ and $a, b \in \{1, 2, 3\}$ such that $a + b \leq 4$, where $\odot$ and $\otimes$ are Hadamard and Kronecker products, i, j denote population index and a, b are exponents.

Assumption 1 essentially replaces normality. Assumption 2 is simple and mild, and as its consequence, $\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)/p^2 = O(1)$. Assumptions 3 and 4 are needed only to control the misclassification rate and consistency of the moments of classifier. Assumption 4 ensures that the moments asymptotically coincide with those under normality whence all terms involving the ratio vanish. The same assumptions will be extended to the multi-sample case in Section 3. The following lemma, proved in Appendix B.1, gives the moments of the classifier. Given $\mathbf{M} > 0, p\times p$, denote $\Delta^2_{\mathbf{M}} = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\mathbf{M}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \mathbf{M}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta^2$ if $\mathbf{M} = \mathbf{I}$.

Lemma 1. *Given the two-sample classifier in Eq. (6), let $\mathbf{x} \in \pi_i$ for some $i \in \{1, 2\}$. Then*

$$\mathrm{E}\{A(\mathbf{x})|\pi_i\} = (-1)^{i+1}\,\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/(2p) = (-1)^{i+1}\Delta^2/(2p), \tag{8}$$

$$\mathrm{var}\{A(\mathbf{x})|\pi_i\} = \delta_i^2/p^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}_i^{-1}}/p^2 = \delta_i^2/p^2 + \Delta^2_{\boldsymbol{\Sigma}_i^{-1}}/p^2, \tag{9}$$

where

$$\delta_i^2 = \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)/n_j + \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\{2n_i(n_i - 1)\}.$$

The moments in Lemma 1 are reported in general notation, $\mathbf{x} \in \pi_i$, so that they are easily extended to the multi-sample case later. Note that the second component in Eq. (9) vanishes under Assumption 3. The rests of $\mathrm{var}\{A(\mathbf{x})|\pi_i\}$, and $\mathrm{E}\{A(\mathbf{x})|\pi_i\}$, are uniformly bounded in p for any fixed n_i, under Assumption 2. Thus, for $p \to \infty$, $\mathrm{E}\{A(\mathbf{x})|\pi_i\}$ and $\mathrm{var}\{A(\mathbf{x})|\pi_i\}$ converge, respectively, to

$$\Delta_0^2/2 \quad \text{and} \quad \delta_{0,i}^2\{O(1/n_1 + 1/n_2) + o(1)\}, \tag{10}$$

where $\lim_{p\to\infty} \Delta^2/p = \Delta_0^2 \in (0, \infty)$ and $\lim_{p\to\infty} \delta_i^2/p^2 = \delta_{0,i}^2 \in (0, \infty)$. The variance vanishes when $n_i, p \to \infty$. It gives consistency of $A(\mathbf{x})$ (see Theorem 2). In practice, consistency must hold with unknown parameters replaced by estimators. We need to estimate Δ^2 and non-vanishing traces in δ_i^2 to estimate the limiting moments of the classifier.

Given $i \in \{1, 2\}$ and $\bar{\mathbf{x}}_i$, $\widehat{\boldsymbol{\Sigma}}_i$ as in (2), let $Q_i = \sum_{k=1}^{n_i} (\widetilde{\mathbf{x}}_{ik}^\top \widetilde{\mathbf{x}}_{ik})^2/(n_i - 1)$, $\widetilde{\mathbf{x}}_i = \mathbf{x}_{ik} - \bar{\mathbf{x}}_i$, and $\eta_i = (n_i - 1)/\{n_i(n_i - 2)(n_i - 3)\}$. The estimators of Δ^2/p^2, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/p^2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)/p^2$ are defined, respectively, as

$$E_0 = U_{n_i} + U_{n_j} - 2U_{n_in_j}, \quad E_i = \eta_i\{(n_i - 1)(n_i - 2)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_i^2) + \{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_i)\}^2 - n_iQ_i\}, \quad E_{ij} = \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{\Sigma}}_j), \tag{11}$$

where U_{n_i} is given after Eq. (6), $U_{n_in_j} = \sum_{k=1}^{n_i}\sum_{\ell=1}^{n_j} A_{ijk\ell}/pn_in_j$ is the corresponding two-sample U-statistic with $\mathrm{E}(U_{n_in_j}) = \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j$, and E_{ij} follows by independence. Note that, E_0, E_i, E_{ij} are also U-statistics based location-invariant estimators. They are presented above in a simplified form which is computationally very efficient. For a detailed treatment in equivalent U-statistic form and proof of their properties under high-dimensional set up, see [1–3]. Theorem 1, proved in Appendix B.2, shows that the variances of the ratios of these estimators to the parameters they estimate are uniformly bounded in p, so that they are consistent for $p \to \infty$.

Theorem 1. *E_0, E_i, E_{ij}, defined in Eq. (11), are unbiased estimators of Δ^2/p^2, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/p^2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)/p^2$. Further, under Assumptions 1–4,*

$$\mathrm{var}(E_0/\Delta^2) = O(1/n_i + 1/n_j), \tag{12}$$
$$\mathrm{var}\{E_i/\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\} = O(1/n_i), \tag{13}$$
$$\mathrm{var}\{E_{ij}/\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)\} = O(1/n_i + 1/n_j), \tag{14}$$
$$\mathrm{cov}\{E_i/\mathrm{tr}(\boldsymbol{\Sigma}_i^2), E_{ij}/\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)\} = O(1/n_i). \tag{15}$$

By Theorem 1, $E_0/\mathrm{E}(E_0) \xrightarrow{\mathcal{P}} 1$ which implies that the empirical mean, $\widehat{\mathrm{E}}\{A(\mathbf{x})|\pi_i\} = E_0/2$, is a consistent estimator of true mean of the classifier. Using similar probability convergence of E_i and E_{ij} implies, by Slutsky's lemma [37, p. 11], that the first component of the empirical variance, $\widehat{\mathrm{var}}\{A(\mathbf{x})|\pi_i\}$, i.e., $\widehat{\delta}_i^2$, is a consistent estimator of δ_i^2 for all $i \in \{1, 2\}$. The limiting empirical moments, parallel to (10), then follow, for $p \to \infty$, as

$$\widehat{\mathrm{E}}\{A(\mathbf{x})|\pi_i\} = \Delta_0^2\{1 + o_P(1)\}/2, \quad \widehat{\mathrm{var}}\{A(\mathbf{x})|\pi_i\} = \delta_{0,i}^2\{O(1/n_1 + 1/n_2) + o_P(1)\}$$

with $\lim_{p\to\infty} E_0/p = \Delta_0^2 \in (0, \infty)$, $\lim_{p\to\infty} \widehat{\delta}_i^2/p^2 = \delta_{0,i}^2 \in (0, \infty)$, where the limit indicates convergence in probability. Hence, with $\mathrm{E}\{\cdot\}$ and $\mathrm{var}\{\cdot\}$ as in Eqs. (8)–(9), for $p \to \infty$,

$$\widehat{\mathrm{E}}\{A(\mathbf{x})|\pi_i\} - \mathrm{E}\{A(\mathbf{x})|\pi_i\} = o_P(1), \tag{16}$$
$$\widehat{\mathrm{var}}\{A(\mathbf{x})|\pi_i\} - \mathrm{var}\{A(\mathbf{x})|\pi_i\} = o_P(1). \tag{17}$$

Theorem 2, proved in Appendix B.3, states the true and empirical consistency of $A(\mathbf{x})$.

Theorem 2. *Consider $A(\mathbf{x})$ in Eq. (6) with its moments as in Lemma 1. For $i \in \{1, 2\}$, let $\mathbf{x} \in \pi_i$. Under Assumptions 1–3, as $n_i, p \to \infty$,*

$$A(\mathbf{x})/(\Delta^2/p) \xrightarrow{\mathcal{P}} (-1)^{i+1}/2 + o_P(1),$$

with Δ^2 defined above. Further, by Eqs. (16)–(17), consistency holds when the moments of the classifier are replaced with their estimates.

By the same arguments, the asymptotic normality of $A(\mathbf{x})$ is given the following theorem, proved in Appendix B.4.

Theorem 3. *Consider $A(\mathbf{x})$ in Eq. (6) with its moments as in Lemma 1. For $i \in \{1, 2\}$, let $\mathbf{x} \in \pi_i$. Under Assumptions 1–3, as $n_i, p \to \infty$,*

$$\frac{A(\mathbf{x}) - \mathrm{E}\{A(\mathbf{x})|\pi_i\}}{\sqrt{\mathrm{var}\{A(\mathbf{x})|\pi_i\}}} \rightsquigarrow \mathcal{N}(0, 1).$$

Further, the limit holds if the moments are replaced with their estimates.

The construction of $A(\mathbf{x})$ is of great benefit in proving Theorem 3. It consists of two parts, each of which is in turn a linear combination of two independent components; this reduces the bulk of the computational burden. Moreover, the optimality property of U-statistics ensures the minimum variance (efficiency) of the classifier. A further verification of these properties through simulations is the subject of Section 4.

2.3. Estimation of misclassification probabilities

Recall the misclassification error $\pi(i|j)$ in Eq. (1) for $i, j \in \{1, 2\}$, $i \neq j$. Given that $\mathrm{E}(\bar{\mathbf{x}}_i) = \boldsymbol{\mu}_i$ and $\mathrm{E}(U_{n_i}) = \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i$ for all $i \in \{1, 2\}$, and assuming the parameters known, consider first the oracle classifier

$$A^{\mathrm{oracle}}(\mathbf{x} \in \pi_1) = \mathbf{x}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/p - (\boldsymbol{\mu}_1^\top\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\mu}_2)/(2p).$$

If $\mathcal{F}_1$ and $\mathcal{F}_2$ are known, say multivariate normal, i.e., $\mathbf{x}_{ik} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, then, assuming $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the so-called optimum error rate of A^{oracle} can be computed as follows, where Φ denotes the standard normal distribution function:

$$\epsilon^{\mathrm{oracle}} = \Phi\left(-\frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{2\sqrt{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}}^2}}\right).$$

Now, assuming equal priors, the best possible performance in this oracle setting, i.e., with $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ known, is achieved by Fisher's linear classifier (equivalently Bayes rule; see [4]), viz.

$$A^{\mathrm{Fisher}}(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2$$

with the corresponding misclassification rate given by

$$\epsilon^{\mathrm{Fisher}} = \Phi\left(-\frac{1}{2}\sqrt{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}}^2}\right),$$

where $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}}$ is the Mahalanobis distance. Denoting ϵ^{Fisher} as a benchmark, the relative performance of $A^{\text{oracle}}(\mathbf{x})$ can be theoretically evaluated by using the ratio of the arguments of Φ, say q, where

$$q = \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\mathbf{I}}}{\{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}} \times \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}}\}^{1/2}}.$$

Bickel and Levina [9] proposed a nice strategy to compute a bound for an expression like q, based on Kantorovich inequality [8]. Following the same idea, let $\mathbf{M}$ be any positive definite symmetric $p \times p$ matrix. Then for any vector $\boldsymbol{\upsilon}$

$$\frac{\|\boldsymbol{\upsilon}\|^2_{\mathbf{I}}}{\|\boldsymbol{\upsilon}\|^2_{\mathbf{M}} \times \|\boldsymbol{\upsilon}\|^2_{\mathbf{M}^{-1}}} \geq \frac{4\lambda_{\min}(\mathbf{M}) \times \lambda_{\max}(\mathbf{M})}{\{\lambda_{\min}(\mathbf{M}) + \lambda_{\max}(\mathbf{M})\}^2},$$

where $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the smallest and the largest eigenvalues of $\mathbf{M}$, respectively. Applying this inequality to q and denoting $\lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma}) = \kappa$ (assuming both eigenvalues bounded away from 0 and ∞), we get

$$q \geq 2\sqrt{\kappa}/(1+\kappa), \tag{18}$$

so that the upper bound on the misclassification probability of $A^{\text{oracle}}(\mathbf{x})$ is

$$\epsilon^{\text{oracle}} \leq \Phi\left\{-\frac{2\sqrt{\kappa}}{1+\kappa}\,\Phi^{-1}(1-\epsilon^{\text{Fisher}})\right\},$$

which essentially depends on κ, the range of non-zero eigenvalues of $\boldsymbol{\Sigma}$. We note that, for moderate κ, the increase in the misclassification rate, induced by taking the covariance matrix away while constructing $A^{\text{oracle}}(\mathbf{x})$, is not large relative to the best possible performance, i.e., ϵ^{Fisher}; see Fig. 4. Further, the upper bound in (18) represents the worst-case scenario so that the empirical results are expected to be better.

Now, for an alternative flavor of the evaluation of the classifier, while still continuing to assume normality, let us condition the classifier on the data, i.e., $A_n(\mathbf{x}) = A(\mathbf{x})|(\overline{\mathbf{x}}_i, U_{n_i})$, say. It immediately follows, for $i \in \{1, 2\}$, that

$$A(\mathbf{x}) \sim \mathcal{N}\left[\boldsymbol{\mu}_i^\top(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)/p - (U_{n_1} - U_{n_2})/2,\ \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2_{\boldsymbol{\Sigma}_i^{-1}}\right].$$

This, using the standardized version of the classifier (Theorem 3), gives the actual error rate, viz.

$$\epsilon_n = \frac{1}{2}\left[\Phi\left\{-\frac{\boldsymbol{\mu}_1^\top(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) - (U_{n_1} - U_{n_2})/2}{\sqrt{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top\boldsymbol{\Sigma}_1(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)}}\right\} + \Phi\left\{-\frac{\boldsymbol{\mu}_2^\top(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) - (U_{n_1} - U_{n_2})/2}{\sqrt{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top\boldsymbol{\Sigma}_2(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)}}\right\}\right],$$

where the subscript n denotes the dependence on the observed sample. Using Theorem 2,

$$\left|\{\boldsymbol{\mu}_i^\top(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)/p - (U_{n_1} - U_{n_2})/2\} - \Delta^2_{0,\mathbf{I}}/2\right| \xrightarrow{\mathcal{P}} 0, \quad \left|\|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2_{\boldsymbol{\Sigma}_i^{-1}} - \Delta^2_{\boldsymbol{\Sigma}_i^{-1}}\right| \xrightarrow{\mathcal{P}} 0,$$

so that by Slutsky's lemma [37, p. 11], as $n_i, p \to \infty$,

$$\epsilon_n - \{\Phi(-\Delta^2_{0,\mathbf{I}}/\Delta_{\boldsymbol{\Sigma}_1^{-1}}) + \Phi(-\Delta^2_{0,\mathbf{I}}/\Delta_{\boldsymbol{\Sigma}_2^{-1}})\}/2 \xrightarrow{\mathcal{P}} 0,$$

where the convergence remains (asymptotically) true even for the sample based classifier $A_n(\mathbf{x})$ since ϵ^{oracle} is the limiting value of ϵ_n.

As the parameters are unknown in practice, they can be replaced with estimates $\widehat{\boldsymbol{\theta}}_j$, leading to empirical regions $\widehat{\mathcal{R}}_i$ which leads to the actual error rate. Although consistent estimators are available, as shown above, the actual rate still cannot be achieved until the form of the underlying distributions is known. As we do not assume any distribution for $A(\mathbf{x})$, normal or otherwise, we resort to the most commonly used practical measure in such situations, namely the apparent error rate, APER, defined as

$$\text{APER} = \sum_{i=1}^{2} \pi_i \widehat{\pi}(i|j) = \frac{m_1 + m_2}{n_1 + n_2}, \tag{19}$$

where $\widehat{\pi}(i|j) = m_i/n_i$ estimates $\pi(i|j)$, m_i is the number of misclassified observations of population i into population j and n_i is the size of the ith sample. Following the standard procedure, we shall combine APER with Lachenbruch's holdout procedure by circulating the training and validation samples; see, e.g., Dudoit et al. [14].

Note that, when we evaluate $A(\mathbf{x})$ by simulating data from an assumed distribution, we can compute the actual or theoretical error rate. We, therefore, use this as a benchmark to compare the estimated APER, using a three-fold cross-validation; see Section 4 for details.

3. The multi-sample case

Here we extend the two-sample classifier to the multi-sample case when $g \geq 2$ populations are independently sampled. For each $i \in \{1, \ldots, g\}$, let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i} \sim \mathcal{F}_i$ be i.i.d vectors with $\mathrm{E}(\mathbf{x}_{ik}) = \boldsymbol{\mu}_i$, $\mathrm{cov}(\mathbf{x}_{ik}) = \boldsymbol{\Sigma}_i$. The multi-sample version of

classifier in Eq. (6) can be expressed, for all $i, j \in \{1, \ldots, g\}$ with $i \neq j$ as

$$A_{ij}(\mathbf{x}) = \mathbf{x}^\top(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)/p - (U_{n_i} - U_{n_j})/2. \tag{20}$$

Alternatively, to write it in a more explicit form, let $A_i(\mathbf{x}) = \mathbf{x}^\top\overline{\mathbf{x}}_i/p - U_{n_i}/2$ be the discriminant function for population i, so that the classifier is

$$A_{ij}(\mathbf{x}) = A_i(\mathbf{x}) - A_j(\mathbf{x}) \tag{21}$$

for any distinct pair (i, j). To extend the classification rule for multi-sample case, let the sample space $\mathcal{X}$ be partitioned into g mutually exclusive regions $\mathcal{R}_1, \ldots, \mathcal{R}_g$ so that, following (1), the error of misclassifying $\mathbf{x}$ to π_j when it actually comes from π_i is given as

$$\pi(j|i) = \Pr(\mathbf{x} \in \mathcal{R}_j|\mathbf{x} \in \pi_i) = \int_{\mathcal{R}_j} d\mathcal{F}_i(\mathbf{x}|\boldsymbol{\theta}).$$

The total probability of misclassification that we aim to minimize, assuming equal costs and priors, is then $p_1 \Pr(1)+\cdots+p_g \Pr(g)$, where $\Pr(i) = \sum_{j=1, j\neq i}^{g} \Pr(j|i)$ denotes the probability of misclassifying an element of π_i into any other population. The classification rule that minimizes this total probability of misclassification can now be defined as

Assign $\mathbf{x}$ to π_i if $\forall_{j\neq i}\, A_{ij}(\mathbf{x}) > 0$, i.e., if $\forall_{j\neq i}\, A_i(\mathbf{x}) > A_j(\mathbf{x})$; otherwise to π_j.

Equivalently, we assign $\mathbf{x}$ to π_i if $A_i(\mathbf{x})$ is the largest among all $i \in \{1, \ldots, g\}$. Since three-sample classification is the most common multi-sample case, the rule can be specifically stated for $g = 3$ as following: Assign $\mathbf{x}$ to π_1 if $A_{12}(\mathbf{x}) > 0$ and $A_{13}(\mathbf{x}) > 0$; assign $\mathbf{x}$ to π_2 if $A_{12}(\mathbf{x}) < 0$ and $A_{23}(\mathbf{x}) > 0$; assign $\mathbf{x}$ to π_3 if $A_{12}(\mathbf{x}) < 0$ and $A_{13}(\mathbf{x}) < 0$.

To study the properties of the multi-sample case, we extend the two-sample assumptions as below.

Assumption 5. $\mathrm{E}(x_{iks}^4) = \gamma < \infty, \gamma \in \mathbb{R}^+$ for all $i \in \{1, \ldots, g\}$.

Assumption 6. $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i)/p = O(1)$ for all $i \in \{1, \ldots, g\}$.

Assumption 7. $\lim_{p\to\infty} \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_l \boldsymbol{\mu}_j/p = O(1)$ for all $i, j, \ell \in \{1, \ldots, g\}$.

Assumption 8. $\lim_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i^a \odot \boldsymbol{\Sigma}_j^b)/\mathrm{tr}(\boldsymbol{\Sigma}_i^a \otimes \boldsymbol{\Sigma}_j^b) = 0$ for all $i, j \in \{1, \ldots, g\}$ and $a, b \in \{1, 2, 3\}$ with $a + b \leq 4$, where $\odot$ and $\otimes$ are the Hadamard and Kronecker products, respectively.

We begin with the following generalization of Lemma 1.

Lemma 2. *Consider $A_{ij}(\mathbf{x})$ in Eq. (20) or (21). Let $\mathbf{x} \in \pi_i$. Then*

$$\mathrm{E}\{A_{ij}(\mathbf{x})|\pi_i\} = (-1)^i\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2/2p = \Delta^2/2p,$$
$$\mathrm{var}\{A_{ij}(\mathbf{x})|\pi_i\} = \delta_i^2/p^2 + \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2_{\boldsymbol{\Sigma}_i^{-1}}/p^2 = \delta_i^2/p^2 + \Delta^2_{\boldsymbol{\Sigma}_i^{-1}}/p^2,$$

where, for all $i, j \in \{1, \ldots, g\}$ with $i \neq j$,

$$\delta_i^2 = \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)/n_j + \sum_{i=1}^{2} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\{2n_i(n_i - 1)\}.$$

The moment estimators follow from the two-sample case in (11) and their consistency from Lemma 1. This helps us extend Theorems 2 and 3 to the general case as follows.

Theorem 4. *Consider $A_{ij}(\mathbf{x})$ in Eq. (20) or (21) with its moments in Lemma 1. For $i \in \{1, \ldots, g\}$, let $\mathbf{x} \in \pi_i$. Under Assumptions 5–7, as $n_i, p \to \infty$,*

$$\frac{A_{ij}(\mathbf{x})}{\Delta^2/p} \xrightarrow{\mathcal{P}} \frac{(-1)^i}{2} + o_P(1), \quad \frac{A_{ij}(\mathbf{x}) - \mathrm{E}\{A_{ij}(\mathbf{x})|\pi_i\}}{\sqrt{\mathrm{var}\{A_{ij}(\mathbf{x})|\pi_i\}}} \rightsquigarrow \mathcal{N}(0, 1).$$

Further, the limit holds if the moments are replaced with their empirical estimators.

As the multi-sample case is a straightforward extension of its two-sample counterpart in Section 2, we skip many detailed proofs to avoid unnecessary repetitions.

4. Simulations

We use simulation results to evaluate the performance of $A(\mathbf{x})$ under practical scenarios, mainly focusing on consistency, asymptotic normality and the control of misclassification under a high-dimensional framework. We consider the case $g = 2$

and generate data from the multivariate normal and Student's t distributions, i.e., $\mathcal{F}_i$ is either $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ or $t_\nu(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $\nu = 10$, $i \in \{1, 2\}$. For each distribution, we set $\boldsymbol{\mu}_1 = \mathbf{0}$ with $\lfloor p/3 \rfloor$ elements of $\boldsymbol{\mu}_2$ also 0 and the rest as 1, where $\lfloor x \rfloor$ denotes the integer part of x. For $\boldsymbol{\Sigma}_i$, we consider two cases:

(1) Both populations have an AR(1) structure, $\text{cov}(x_k, x_\ell) = \sigma^2 \rho^{|k-\ell|}$ for all k, ℓ, with $\sigma^2 = 1$ for $i = 1$ and 2, where $\rho = 0.3$ for $i = 1$ and $\rho = 0.7$ for $i = 2$, to represent both low and high correlation structures.
(2) Same AR(1) structure for $i = 1$ with $\sigma^2 = 1$, $\rho = 0.5$, whereas unstructured (UN) $\boldsymbol{\Sigma}_i$ for $i = 2$, defined as $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^{p}$ with $\sigma_{ii} = 1(1)p$, $\rho_{ij} = (i-1)/p$, $i \neq j$.

For finite-sample performance of the classifier under growing dimension, emphasizing $p \gg n_i$, we draw samples of sizes $n_1 = 5, n_2 = 7$, with $p = \{10, 20, 50, 100, 300, 500, 700, 1000, 3000, 5000, 10000\}$. Finally, all results are averages of 1000 simulation runs for each combination of parameters mentioned above. Additionally, to observe the effect of large n_i, the misclassification rates are also presented for $n_1 = 10$, $n_2 = 12$. We also assessed the classifier for very different sample sizes, e.g., $n_1 = 5, n_2 = 25$ or $n_1 = 10, n_2 = 50$, with similar results, hence not reported here.

Fig. 1 shows the QQ-plots for asymptotic normality of $A(\mathbf{x})$, whose first two rows are for the normal distribution, respectively for AR–AR and AR–UN structures; likewise, the last two rows are for Student's t distribution with $\nu = 10$. The three QQ-plots in each row pertain to $p \in \{100, 500, 1000\}$ dimensions (left to right). As stated above, similar results were obtained for other dimensions, up to $p = 10{,}000$, but only a selection is reported here.

We observe a very close normal approximation for n_i as small as 5 or 7, and the results for Student's t distribution depict small-sample robustness of the classifier to non-normality. Comparing the results for two distributions, the heavy-tailed behavior of Student's t distribution translates into a small departure of points from the line at the extremes. But in general, a nice normal approximation holds for both distributions, and is not altered as the dimension grows.

A similar performance is observed for the control of misclassification rate, shown in Figs. 2 for $n_1 = 5$, $n_2 = 7$, and 3 for $n_1 = 10$, $n_2 = 12$. The thick line represents the actual error rate under asymptotic normality of the classifier, i.e., $\Phi\{-\text{E}(A)/\sqrt{\text{var}(A)}\}$, where Φ is the (univariate) standard normal distribution. This actual error rate is used as a reference to assess the estimated error rate shown in the dashed line for the normal distribution, and in the dotted line for the t_{10} distribution. Further, the upper and lower panels in each figure are for the AR–AR and AR–UN pair of covariances, respectively.

The estimated error closely follows the actual error for $n_1 = 5, n_2 = 7$, and the error rate also converges to zero, showing consistency of the classifier. For Student's t distribution with $n_1 = 5, n_2 = 7$, the estimated error rates are relatively higher than under normality, but with n_i increased only by 5, a discernible difference in the performance of the classifier is observed in Fig. 3. Note that the x-axis in Figs. 2–3 is truncated at $p = 500$ since the misclassification rates already converge to 0 by this value and remain so for larger p.

5. Applications

We apply $A(\mathbf{x})$ on two large data sets for $g \in \{2, 3\}$. With moderate sample sizes (77 and 102), we use $K = 3$-fold CV for evaluation; see [14]. Let $\mathcal{L}$ and $\mathcal{T}$ be learning and test sets. We randomly divide the data set into K classes of roughly equal size, where $\mathcal{T}$ consists of $K - 1$ classes and the Kth class held out as test data. The procedure is repeated K times, each time with a different test class, and a misclassification rate is computed for each repetition. The evaluation criterion is the average misclassification rate over all repetitions.

For the kth fold of CV, let $n_i^k(\mathcal{L})$, $n_i^k(\mathcal{T})$ and $m_{ij}^k(\mathcal{T})$ be, respectively, the sample sizes for learning and test data in sample i and the number of misclassified observations from class i into class j, $i, j \in \{1, \ldots, g\}$, $g = 2$ or 3, $k \in \{1, \ldots, K\}$, $K = 3$. Let $e^k(i|j)$ be the estimated misclassification rate, an estimator of $\pi(i|j)$ in (1), for kth rotation, i.e., $e^k(i|j) = m_{ij}^k(\mathcal{T})/n^k(\mathcal{T})$, where $n^k(\mathcal{T}) = n_i^k(\mathcal{T}) + n_j^k(\mathcal{T})$. For $g = 3$, we do the same procedure for each of three pairs and compute overall misclassification rate. For details, see [7,35,36] etc.

Example 1 (*DLBCL Data*)**.** The Diffuse Large B-cell Lymphoma (DLBCL) data belongs to a study of lymphoid malignancy in adults. The analysis reported here consists of $p = 5469$ gene expressions studied on pre-treatment biopsies from two independent groups of 77 patients, one with DLBCL ($n_1 = 58$), the other with follicular lymphoma (FL) ($n_2 = 19$). For a 3-fold CV, we randomly divide the data into three groups of sizes 26, 26, 25 with $n^1(\mathcal{L}) = 52$, $n^1(\mathcal{T}) = 25$, $n^k(\mathcal{L}) = 51$, $n^k(\mathcal{T}) = 26$ for $k \in \{2, 3\}$. By coding the populations as 1 (DLBCL) and 2 (FL), the misclassifications observed from the three rotations of CV, i.e., m_{12}^k and m_{21}^k, are

$$m_{12}^1 = 3,\ m_{21}^1 = 1; \quad m_{12}^2 = 6,\ m_{21}^2 = 0; \quad m_{12}^3 = 2,\ m_{21}^3 = 3,$$

with an overall misclassification rate 15/77. The sample sizes for each fold are as below.

$$K = 1 : n_1^1(\mathcal{L}) = 38,\ n_2^1(\mathcal{L}) = 14,\ n_1^1(\mathcal{T}) = 20,\ n_2^1(\mathcal{T}) = 5,$$
$$K = 2 : n_1^2(\mathcal{L}) = 40,\ n_2^2(\mathcal{L}) = 11,\ n_1^2(\mathcal{T}) = 18,\ n_2^2(\mathcal{T}) = 8,$$
$$K = 3 : n_1^3(\mathcal{L}) = 38,\ n_2^3(\mathcal{L}) = 13,\ n_1^3(\mathcal{T}) = 20,\ n_2^3(\mathcal{T}) = 6$$

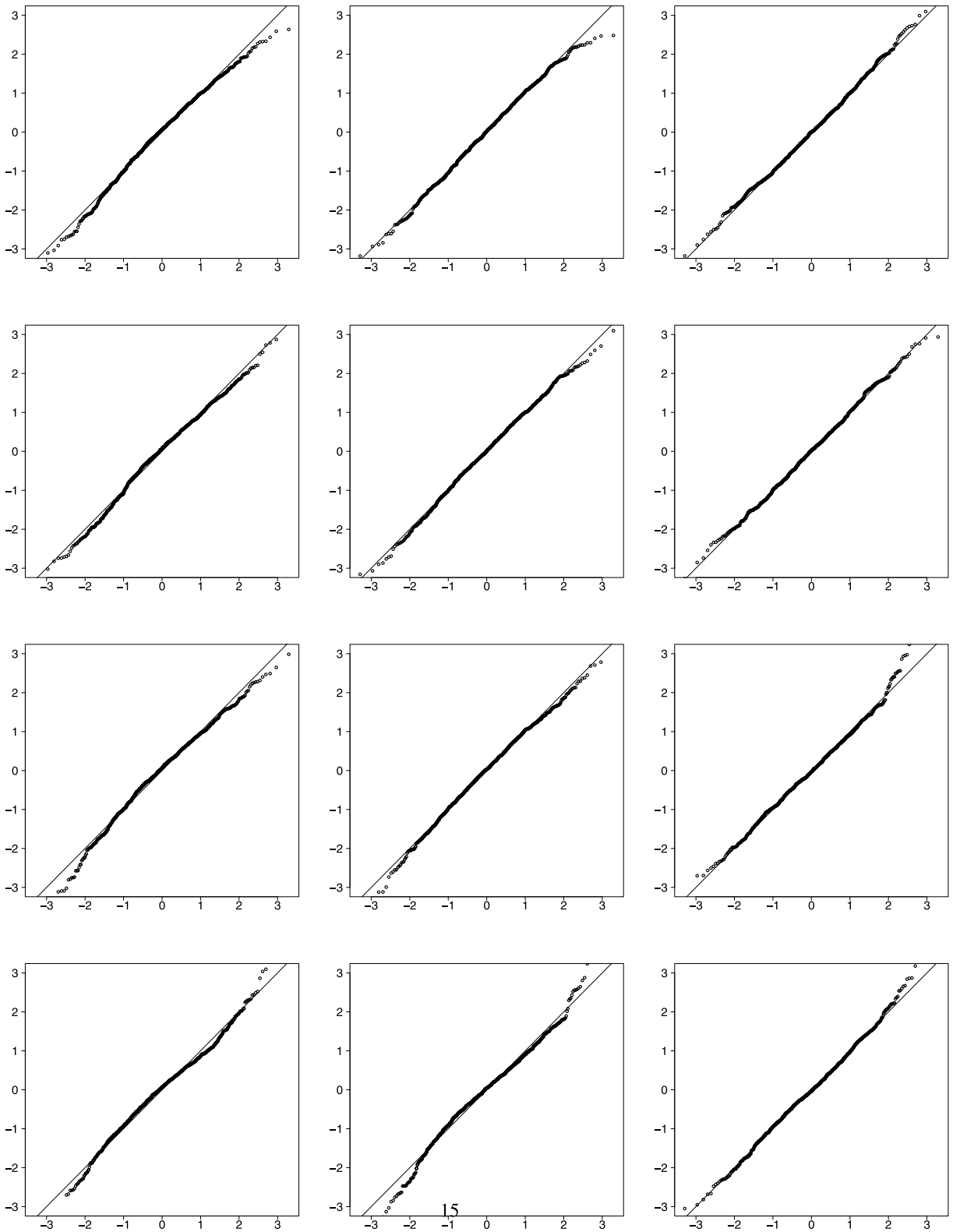

Fig. 1. QQ plots of $A(\mathbf{x})$ for two-class case for multivariate normal (Rows 1–2) and t (Rows 3–4) distributions, $n_1 = 5$, $n_2 = 7$, $p \in \{100, 500, 1000\}$ (L-R each row) and covariance structures AR–UN (Rows 1 and 3) and AR–AR (Rows 2 and 4).

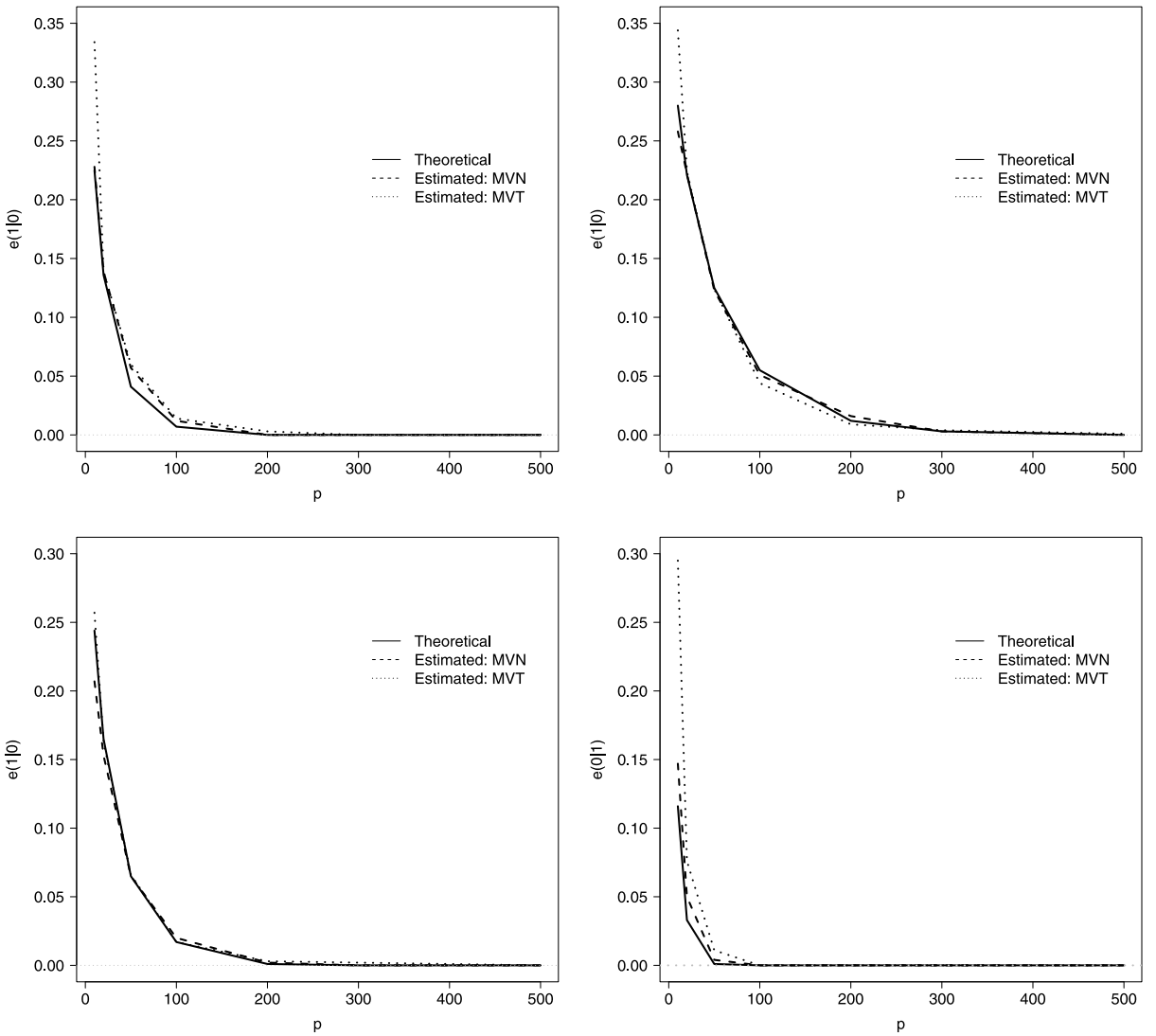

Fig. 2. Theoretical (thick line) and estimated error rates (APER, Eq. (19)) of $A(\mathbf{x})$ for two-class case for multivariate normal (dashed) and t (dotted) distributions, $n_1 = 5$, $n_2 = 7$, $p \in \{10, 20, 50, 100, 200, 300, 500\}$, covariance structures AR–AR (upper panel) and AR–UN (lower penal).

Example 2 (*Leukemia Data*)**.** The data set pertains to a study of patients with acute lymphoblastic leukemia (ALL) carrying a chromosomal translocation involving mixed-lineage leukemia (MLL) gene. The analysis reported here consists of $p = 11{,}225$ gene expression profiles of leukemia cells from $n_2 = 24$ patients diagnosed with B-precursor ALL carrying an MLL translocation and compared to a group of $n_3 = 20$ individual diagnosed with conventional B-precursor without MLL translocation. In addition, there is a third group of a random sample of $n_1 = 28$ with acute myelogenous leukemia (AML).

For a 3-fold cross-validation, we randomly divide the data into three equal groups of size 24 and use $K - 1 = 2$ classes of total $n^k(\mathcal{L}) = 48$ observations in learning set and $n^k(\mathcal{T}) = 24$ in the test set, $k \in \{1, 2, 3\}$. The rest of the procedure is same as in Example 1. The misclassifications, given below, lead to an overall misclassification rate 9/72.

$$
\begin{aligned}
&K = 1 : m^1_{12} = 1,\ m^1_{21} = 0; \quad m^1_{13} = 2,\ m^1_{31} = 1; \quad m^1_{23} = 0,\ m^1_{32} = 1,\\
&K = 2 : m^2_{12} = 0,\ m^2_{21} = 1; \quad m^2_{13} = 2,\ m^2_{31} = 0; \quad m^2_{23} = 1,\ m^2_{32} = 0,\\
&K = 3 : m^3_{12} = 0,\ m^3_{21} = 0; \quad m^3_{13} = 0,\ m^3_{31} = 0; \quad m^3_{23} = 0,\ m^3_{32} = 0.
\end{aligned}
$$

The sample sizes used in each rotation are as follows:

$$
\begin{aligned}
&K = 1 : n^1_1(\mathcal{L}) = 21,\ n^1_2(\mathcal{L}) = 15,\ n^1_3(\mathcal{L}) = 12; \quad n^1_1(\mathcal{T}) = 7,\ n^1_2(\mathcal{T}) = 9,\ n^1_3(\mathcal{T}) = 8,\\
&K = 2 : n^2_1(\mathcal{L}) = 18,\ n^2_2(\mathcal{L}) = 14,\ n^2_3(\mathcal{L}) = 16; \quad n^2_1(\mathcal{T}) = 10,\ n^2_2(\mathcal{T}) = 10,\ n^2_3(\mathcal{T}) = 4,\\
&K = 3 : n^3_1(\mathcal{L}) = 17,\ n^3_2(\mathcal{L}) = 19,\ n^3_3(\mathcal{L}) = 12; \quad n^3_1(\mathcal{T}) = 11,\ n^3_2(\mathcal{T}) = 5,\ n^3_3(\mathcal{T}) = 8.
\end{aligned}
$$

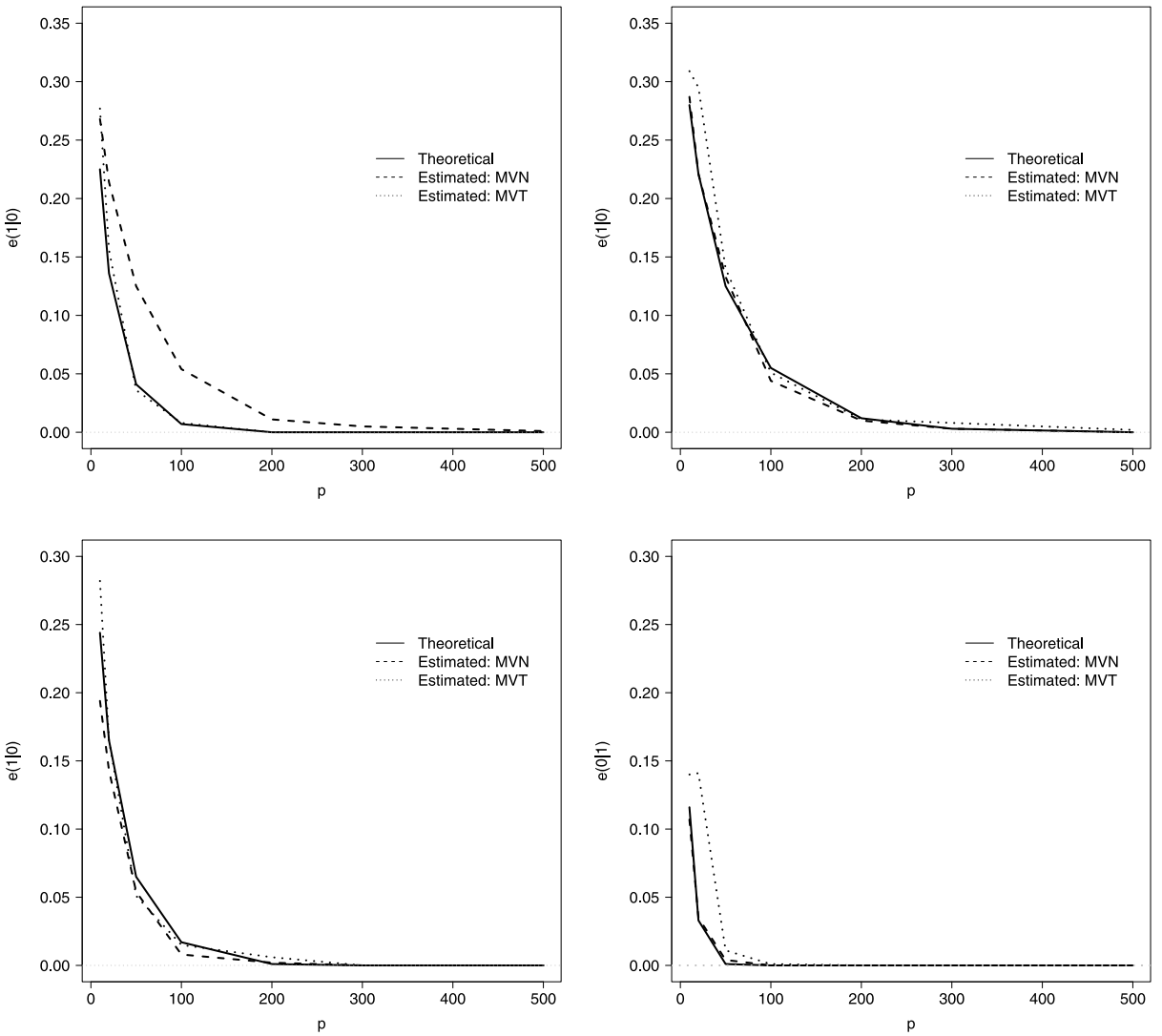

Fig. 3. Theoretical (thick line) and estimated error rates (APER, Eq. (19)) of $A(\mathbf{x})$ for two-class case for multivariate normal (dashed) and t (dotted) distributions, $n_1 = 10, n_2 = 12, p \in \{10, 20, 50, 100, 200, 300, 500\}$, covariance structures AR–AR (upper panel) and AR–UN (lower penal).

6. Discussion and conclusions

A U-classifier for high-dimensional and possibly non-normal data has been proposed. The threshold part of the classifier, called U-component, is a linear combination of two bivariate U-statistics computed from two independent samples. The discriminator, called P-component, forms an inner product between the observation to be classified and the difference of the mean vectors of the corresponding independent samples. It results into a computationally simple classifier which is linear without requiring the underlying covariance matrices to be equal. A multi-class extension with same properties has also been given.

The classifier is unbiased, consistent and asymptotically normal. Rapid convergence of error rate has been shown for small sample sizes and non-normal distributions, under mild and practical conditions. The performance of the classifier, in terms of its consistency, asymptotic normality and control of misclassification rate, has been shown through simulations for different distributions with small sample sizes and large dimension.

We applied the classifier to genetics and microarray data sets, some of the most popular areas for classification analysis. To emphasize the role of high-dimensionality, we demonstrated that the use, accuracy, and validity of the classifier does not rest on any form of data pre-processing. That is, a data set measured in large dimension can be directly used for classification without any dimension reduction through sorting, clustering or other means.

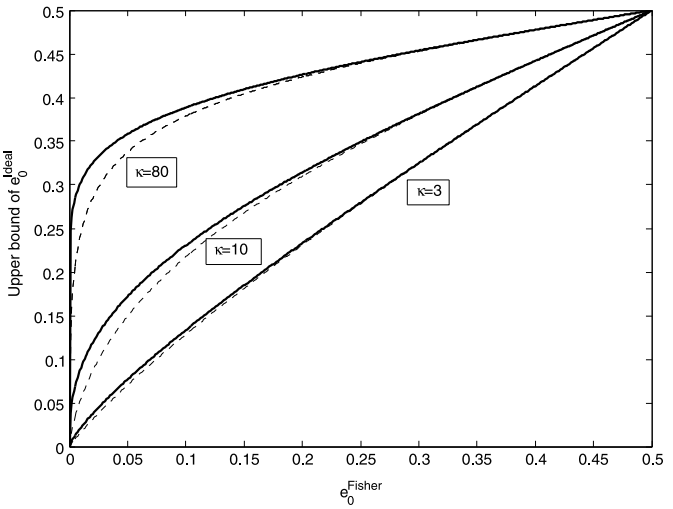

Fig. 4. Upper bound on the misclassification probability of $A^{\text{oracle}}(\mathbf{x})$ as a function of $\epsilon_o^{\text{Fisher}}$ for normal (thick line) and t_5 (dashed line) distributions with $\kappa \in \{3, 10, 80\}$.

Acknowledgments

The authors would like to thank the Editor-in-Chief, Christian Genest, and the anonymous reviewers for their constructive comments and suggestions on the original version of this paper. The second author is supported by Grant 2013-45266 of the National Research Council of Sweden (VR).

Appendix A. Moments of quadratic and bilinear forms

For $\mathbf{z}_{ik}$ in (7), let $A_{ik} = \mathbf{z}_{ik}^\top \Sigma_i \mathbf{z}_{ik} = \mathbf{y}_{ik}^\top \mathbf{A}^2 \mathbf{y}_{ik}$, $A_{ijk\ell} = \mathbf{z}_{ik}^\top \mathbf{z}_{j\ell} = \mathbf{y}_{ik}^\top \mathbf{\Lambda}_i \mathbf{\Lambda}_j \mathbf{y}_{j\ell}$, $k \neq \ell$, be quadratic and bilinear forms of independent components with $A_{ik} = \mathbf{y}_{ik}^\top \mathbf{y}_{ik} = Q_{ik}$ if $\Sigma_i = \mathbf{I}$. As all terms involving $\mathbf{A}_i$ eventually vanish under Assumption 4, we write $\mathbf{A}_i = \mathbf{A}$ for simplicity. Theorem 5 gives basic moments of A_{ik} and $A_{ijk\ell}$, which are extended in Lemma 3. Proofs of these results are tedious but simple, hence skipped; see [2].

Theorem 5. *For A_{ik} and $A_{ijk\ell}$, as defined above, we have*

$$\mathrm{E}(Q_{ik}^2)^2 = 2\,\mathrm{tr}(\Sigma_i^2) + \{\mathrm{tr}(\Sigma_i)\}^2 + M_1, \quad \mathrm{E}(A_{ik}^2)^2 = 2\,\mathrm{tr}(\Sigma_i^4) + \{\mathrm{tr}(\Sigma_i^2)\}^2 + M_2,$$
$$\mathrm{E}(A_{ijk\ell}^4) = 6\,\mathrm{tr}(\Sigma_i\Sigma_j)^2 + 3\{\mathrm{tr}(\Sigma_i\Sigma_j)\}^2 + M_3,$$

$$\mathrm{E}(A_{ik}A_{jk}) = 2\,\mathrm{tr}(\Sigma_i^3\Sigma_j) + \mathrm{tr}(\Sigma_i^2)\,\mathrm{tr}(\Sigma_i\Sigma_j) + M_2,$$
$$\mathrm{E}(Q_{ik}Q_{jk}A_{ijk\ell}^2) = 4\,\mathrm{tr}(\Sigma_i\Sigma_j)^2 + 4\,\mathrm{tr}(\Sigma_i^3)\,\mathrm{tr}(\Sigma_j) + \{\mathrm{tr}(\Sigma_i)\}^2\,\mathrm{tr}(\Sigma_j^2) + M_4,$$

with $M_1 = \gamma\,\mathrm{tr}(\mathbf{A} \odot \mathbf{A})$, $M_2 = \gamma\,\mathrm{tr}(\mathbf{A}^2 \odot \mathbf{A}^2)$,

$$M_3 = 6\gamma\,\mathrm{tr}(\mathbf{A}^2 \odot \mathbf{A}^2) + \gamma^2 \sum_{s=1}^{p}\sum_{t=1}^{p} A_{st}^4, \quad M_4 = 2\gamma\,\mathrm{tr}(\Sigma_i)\,\mathrm{tr}(\mathbf{A}^2 \odot \mathbf{A}) + 4\gamma\,\mathrm{tr}(\mathbf{A}^3 \odot \mathbf{A}) + \gamma\,\mathrm{tr}(\mathbf{A} \odot \mathbf{ADA}),$$

and $\mathbf{D} = \mathrm{diag}(\mathbf{A})$. *Moreover,* $\mathrm{E}(A_{ik}) = \mathrm{tr}(\Sigma_i)$, $\mathrm{E}(A_{ikr}^2) = \mathrm{tr}(\Sigma_i^2)$ *and* $\mathrm{cov}(A_{ik}, A_{ikr}) = 0$.

Lemma 3. *Let $\mathbf{z}_{it}$ be as given above with $\mathbf{z}_{it}$, $\mathbf{z}_{iu}$ independent if $t \neq u$. Then*

$$\mathrm{E}(\mathbf{z}_{it}^\top\mathbf{z}_{iu}\mathbf{z}_{it}^\top\mathbf{z}_{iv}\mathbf{z}_{iu}^\top\Sigma_i\mathbf{z}_{iv}) = \mathrm{tr}(\Sigma_i^4) = \mathrm{E}\{\mathbf{z}_{it}^\top\mathbf{z}_{iu}\mathbf{z}_{iw}^\top\mathbf{z}_{iu}\mathbf{z}_{it}^\top\mathbf{z}_{iv}\mathbf{z}_{iw}^\top\mathbf{z}_{iv}t\}, \quad \mathrm{cov}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2, \mathbf{z}_{it}^\top\Sigma_j\mathbf{z}_{it}\} = 2\,\mathrm{tr}(\Sigma_i^3\Sigma_j) + M_2,$$
$$\mathrm{E}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2\mathbf{z}_{it}^\top\Sigma_i\mathbf{z}_{it}\} = 2\,\mathrm{tr}(\Sigma_i^4) + \{\mathrm{tr}(\Sigma_i^2)\}^2 + M_2 = \mathrm{var}(\mathbf{z}_{it}^\top\mathbf{z}_{iu}\mathbf{z}_{iv}^\top\mathbf{z}_{iu}), \quad \mathrm{E}\{(\mathbf{z}_{iu}^\top\mathbf{z}_{iv})^2\mathbf{z}_{it}^\top\Sigma_j\mathbf{z}_{it}\} = \mathrm{tr}(\Sigma_i\Sigma_j)\,\mathrm{tr}(\Sigma_i)^2,$$
$$\mathrm{E}\{(\mathbf{z}_{it}^\top\mathbf{z}_{ju})^2\mathbf{z}_{it}^\top\Sigma_j\mathbf{z}_{it}\} = 2\,\mathrm{tr}\{(\Sigma_i\Sigma_j)^2\} + \{\mathrm{tr}(\Sigma_i\Sigma_j)\}^2 + M_2 = \mathrm{var}(\mathbf{z}_{it}^\top\mathbf{z}_{ju}\mathbf{z}_{iv}^\top\mathbf{z}_{ju}), \quad \mathrm{cov}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2, (\mathbf{z}_{it}^\top\mathbf{z}_{iv})^2\} = 2\,\mathrm{tr}(\Sigma_i^4) + M_2,$$
$$\mathrm{E}(\mathbf{z}_{jt}^\top\mathbf{z}_{iu}\mathbf{z}_{jt}^\top\mathbf{z}_{iv}\mathbf{z}_{iu}^\top\Sigma_j\mathbf{z}_{iv}) = \mathrm{tr}\{(\Sigma_i\Sigma_j)^2\} = \mathrm{cov}(\mathbf{z}_{it}^\top\Sigma_i\mathbf{z}_{iu}, \mathbf{z}_{it}^\top\Sigma_j\mathbf{z}_{iu}), \quad \mathrm{cov}\{(\mathbf{z}_{jt}^\top\mathbf{z}_{iu})^2, (\mathbf{z}_{jt}^\top\mathbf{z}_{iv})^2\} = 2\,\mathrm{tr}\{(\Sigma_i\Sigma_j)^2\} + M_2.$$

Furthermore, $\mathrm{E}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2\mathbf{z}_{it}^\top\mathbf{z}_{iu}\mathbf{z}_{it}^\top\mathbf{z}_{iv}\}$, $\mathrm{E}\{\mathbf{z}_{it}^\top\mathbf{z}_{iu}\mathbf{z}_{it}^\top\mathbf{z}_{iv}\mathbf{z}_{it}^\top\Sigma_i\mathbf{z}_{it}\}$, $\mathrm{E}\{\mathbf{z}_{it}^\top\mathbf{z}_{iu}\mathbf{z}_{it}^\top\mathbf{z}_{iv}\mathbf{z}_{it}^\top\Sigma_i\mathbf{z}_{iu}\}$, $\mathrm{E}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2\mathbf{z}_{it}^\top\Sigma_i\mathbf{z}_{iu}\}$, $\mathrm{E}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2\mathbf{z}_{it}^\top\Sigma_j\mathbf{z}_{iu}\}$, *and* $\mathrm{E}\{(\mathbf{z}_{it}^\top\mathbf{z}_{iu})^2\mathbf{z}_{it}^\top\mathbf{z}_{iv}\mathbf{z}_{iu}^\top\mathbf{z}_{iv}\}$ *all vanish.*

Appendix B. Main proofs

In what follows, $\mathrm{E}(\cdot)$ and $\mathrm{var}(\cdot)$ will denote, for simplicity, $\mathrm{E}(\cdot|\pi_1)$ and $\mathrm{var}(\cdot|\pi_1)$.

B.1. Proof of Lemma 1

Let $\mathbf{x} \in \pi_1$. With $\mathbf{x}$ independent of both samples, $\mathrm{E}\{A(\mathbf{x})\}$ is trivial. For variance, ignoring p for simplicity, we begin with

$$\mathrm{var}\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\} = \mathrm{E}\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\}^2 - \{\boldsymbol{\mu}_1^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}^2.$$

Since $\mathrm{E}\left\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right\}^2 = \mathrm{E}\left\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top\mathbf{x}\right\}$, we immediately get

$$\mathrm{E}\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\}^2 = \mathrm{tr}[(\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top)\{\boldsymbol{\Sigma}_1/n_1 + \boldsymbol{\Sigma}_2/n_2 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\}],$$

so that

$$\mathrm{var}\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\} = \mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2 + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1\boldsymbol{\mu}_1/n_1 + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_2\boldsymbol{\mu}_1/n_2 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Now $\mathrm{var}(U_{n_1} - U_{n_2}) = \mathrm{var}(U_{n_1}) + \mathrm{var}(U_{n_2})$. For U_{n_i} with $h(\mathbf{x}_{ik}, \mathbf{x}_{ir}) = \mathbf{x}_{ik}^\top\mathbf{x}_{ir}$, $h_1(\mathbf{x}_{ik}) = \mathbf{x}_{ik}^\top\boldsymbol{\mu}_i$ with $\xi_1 = \mathrm{var}\{h_1(\mathbf{x}_{ik})\} = \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i$, and $h_2 = h$ with $\xi_2 = \mathrm{var}(A_{ik}) = \mathrm{tr}(\boldsymbol{\Sigma}_i^2) + 2\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i$, so that, for $i \in \{1, 2\}$,

$$\mathrm{var}(U_{n_i}) = 2\{2(n_i - 2)\xi_1 + \xi_2\}/\{n_i(n_i - 1)\} = 2\,\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\{n_i(n_i - 1)\} + 4\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i/n_i.$$

For $\mathrm{cov}\left\{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), U_{n_1} - U_{n_2}\right\}$, $\mathrm{cov}(\mathbf{x}^\top\bar{\mathbf{x}}_2, U_{n_1}) = 0 = \mathrm{cov}(\mathbf{x}^\top\bar{\mathbf{x}}_1, U_{n_2})$, by independence, where it immediately follows that $\mathrm{cov}(\mathbf{x}^\top\bar{\mathbf{x}}_i, U_{n_i}) = 2\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i/n_i$, $i \in \{1, 2\}$. Combining all results and simplifying gives Eq. (9). □

B.2. Proof of Theorem 1

The unbiasedness is trivial. For $\mathrm{var}(E)$, $\mathrm{var}(U_{n_i})$ for each $i \in \{1, 2\}$ is given in Appendix B.1. For $\mathrm{var}(U_{n_{ij}})$, $h(\mathbf{x}_{ik}, \mathbf{x}_{j\ell}) = \mathbf{x}_{ik}^\top\mathbf{x}_{j\ell}$ with $h_{10} = \boldsymbol{\mu}_j^\top\mathbf{x}_{ik}$, $h_{01} = \boldsymbol{\mu}_i^\top\mathbf{x}_{j\ell}$ so that $\xi_{10} = \mathrm{var}\{h_{10}(\cdot)\} = \boldsymbol{\mu}_j^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_j$ and $\xi_{10} = \mathrm{var}\{h_{10}(\cdot)\} = \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_j\boldsymbol{\mu}_i$. Also $h_{11}(\cdot) = h(\cdot)$ with $\xi_{11} = \mathrm{var}\{h_{11}(\cdot)\} = \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_j\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_j + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)$. Hence, from [24],

$$\mathrm{var}(U_{n_in_j}) = \frac{1}{n_in_jp^2}\{n_i\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_j\boldsymbol{\mu}_i + n_j\boldsymbol{\mu}_j^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_j + \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)\}$$

where $\mathrm{cov}(U_{n_i}, U_{n_in_j}) = 2\boldsymbol{\mu}_j^\top\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i/n_ip^2$, $\mathrm{cov}(U_{n_j}, U_{n_in_j}) = 2\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}_j\boldsymbol{\mu}_j/n_jp^2$ and, by independence, $\mathrm{cov}(U_{n_i}, U_{n_j}) = 0$. Now $\mathrm{var}(E/p)$ can be approximated as

$$\mathrm{var}(E/p) = 2\,\mathrm{tr}(\boldsymbol{\Sigma}_{0ij}^2)/p^2 + 4(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top\boldsymbol{\Sigma}_{0ij}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/p^2,$$

for $i, j \in \{1, 2\}$, $i \neq j$, where $\boldsymbol{\Sigma}_{0ij} = \boldsymbol{\Sigma}_i/n_i + \boldsymbol{\Sigma}_j/n_j$. The second term vanishes under Assumption 3 and the first is bounded in p under Assumption 2 so that $\mathrm{var}(E_0)$ reduces to $O(1/n_1 + 1/n_2)$ as $p \to \infty$, providing consistency. The bound in (12) follows trivially. As E_i and E_{ij}, are one- and two-sample U-statistics with higher order kernels, we essentially follow the same strategy as for E. First, from Theorem 5 and Lemma 3, it can be shown that

$$\begin{aligned}
\mathrm{var}(E_i) &= \frac{4}{\eta(n_i)p^4}\left[(2n_i^3 - 9n_i^2 + 9n_i - 16)\,\mathrm{tr}(\boldsymbol{\Sigma}_i^4) + (n_i^2 - 3n_i + 8)\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}^2 + M_2O(n_i^3) + M_3O(n_i^2)\right],\\
\mathrm{var}(E_{ij}) &= \frac{2}{(n_i - 1)(n_j - 1)p^4}\left\{(n - 1)\,\mathrm{tr}\{(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)^2\} + \{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2 + M_2O(n) + M_3O(1)\right\},\\
\mathrm{cov}(E_i, E_{ij}) &= \frac{4}{n_i(n_i - 1)p^4}\{n_i\,\mathrm{tr}(\boldsymbol{\Sigma}_i^3\boldsymbol{\Sigma}_j) + M_2O(n_i)\},
\end{aligned}$$

with $n = n_i + n_j$, $i, j \in \{1, 2\}$, $i \neq j$, $\eta(n_i) = n_i(n_i - 1)(n_i - 2)(n_i - 3)$, M_2, M_3 are as in Theorem 5 and $\mathrm{cov}(E_i, E_j) = 0$. As the terms involving Ms vanish under Assumption 4, the consistency and the bounds in Eqs. (13)–(15) follow, by Cauchy–Schwarz inequality, the same way as for E_0. Note also that, the terms involving M's are exactly zero under normality whence the same results hold even more conveniently. □

B.3. Proof of Theorem 2

The proof essentially follows from that of Theorem 1 without much new computations. In particular, the first part, assuming true parameters known, is trivial. For the second part with estimates, the (n_i, p)-consistency of estimators proved in Appendix B.2 implies that $E/\mathrm{E}(E) \xrightarrow{\mathcal{P}} 1$, and the same holds for E_i, E_{ij}.

Plugging these estimators in the moments of $A(\mathbf{x})$ and using Slutsky's lemma, $\widehat{\delta}_i^2/\delta_i^2 \xrightarrow{\mathcal{P}} 1$ so that $\widehat{\mathrm{var}}\{A(\mathbf{x})\} = \mathrm{var}\{A(\mathbf{x})\} + o_P(1)$, and the consistency follows similarly as with known parameters. □

B.4. Proof of Theorem 3

Write $\widetilde{A}(\mathbf{x}) = \{A(\mathbf{x})|\mathbf{x} \in \pi_1\} - \mathrm{E}\{A(\mathbf{x})|\mathbf{x} \in \pi_1\}$, where, ignoring p for simplicity,

$$\widetilde{A}(\mathbf{x}) = \{\mathbf{x}^\top(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \boldsymbol{\mu}_1^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\} - \{(U_{n_1} - \boldsymbol{\mu}_1^\top\boldsymbol{\mu}_1) - (U_{n_2} - \boldsymbol{\mu}_2^\top\boldsymbol{\mu}_2)\}/2,$$

Let $\widehat{U}_{n_i}$ be the projection of $\widetilde{U}_{n_i} = U_{n_i} - \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i$. Then $g_1(\mathbf{x}_{1k}) = h_1(\cdot) - \boldsymbol{\mu}_1^\top\boldsymbol{\mu}_1 = (\mathbf{x}_{1k} - \boldsymbol{\mu}_1)^\top\boldsymbol{\mu}_1$ for U_{n_1}; see Chapter 5 in [33]. Similarly, $g_1(\mathbf{x}_{2\ell})$ for U_{n_2}, with $\mathrm{E}\{g_1(\cdot)\} = 0$ for both. Thus

$$\widehat{U}_{n_1} - \widehat{U}_{n_2} = \frac{2}{n_1}\sum_{k=1}^{n_1}(\mathbf{x}_{1k} - \boldsymbol{\mu}_1)^\top\boldsymbol{\mu}_1 - \frac{2}{n_2}\sum_{\ell=1}^{n_2}(\mathbf{x}_{2\ell} - \boldsymbol{\mu}_2)^\top\boldsymbol{\mu}_2, \tag{B.1}$$

where $\widetilde{U}_{n_i} = \widehat{U}_{n_i} + o_P(1)$ for all $i \in \{1, 2\}$. With $\mathbf{x} \in \pi_1$ and independence of samples, this projection of $\widetilde{A}(\mathbf{x})$ results into a sum of two independent components, each an average of i.i.d variables [37]. For fixed p, the normal limit follows immediately by the usual Central Limit Theorem. However, the kernels of U-statistics here vary with n_i (hence with p through n_i).

The theory of U-statistics with varying kernel is rich and has been considered by many authors, e.g., [22,29]; see also Chapter 4 in [23] and the references therein. In fact, the theory has recently also got inroads into high-dimensional statistics. An approach with very similarly constructed U-statistics, as in our case, is given in [30,40], where the latter also provide a general framework for the application of U-statistics in a high-dimensional set up. See also [2].

The key factor in determining the limit of such U-statistics rests on the projection variances. This limit, in our case, follows relatively easily for two reasons. First, $A(\mathbf{x})$, hence its projection in Eq. (B.1), is a sum of two independent components. Second, the projection variances are uniformly bounded under the assumptions so that the variance of the linear combination of these projections in Eq. (B.1), and eventually the variance of the classifier itself, is uniformly bounded. The limit then follows immediately from Theorem 6.1.2 in [25]. □

References

[1] M.R. Ahmad, A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens–Fisher setting, Ann. Inst. Statist. Math. 66 (2014) 33–61.
[2] M.R. Ahmad, Location-invariant multi-sample U-tests for covariance matrices with large dimension, Scand. J. Stat. 44 (2017) 500–523.
[3] M.R. Ahmad, Location-invariant tests of homogeneity of large dimensional covariance matrices, J. Stat. Theory Pract. 11 (2017) 731–745.
[4] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, third ed. Wiley, New York, 2003.
[5] M. Aoshima, K. Yata, A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, Ann. Inst. Statist. Math. 66 (2014) 983–1010.
[6] M. Aoshima, K. Yata, Geometric classifier for multiclass, high-dimensional data, Sequential Anal. 34 (2015) 279–294.
[7] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, S.J. Korsmeyer, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nature Genet. 30 (2002) 41–47.
[8] D.S. Bernstein, Matrix Mathematics: Theory, Facts and Formulas, second ed. Princeton University Press, Princeton, NJ, 2009.
[9] P. Bickel, E. Levina, Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, Bernoulli 10 (2004) 989–1010.
[10] T. Cai, W. Liu, A direct estimation approach to sparse linear discriminant analysis, J. Amer. Statist. Assoc. 106 (2011) 1566–1577.
[11] Y.-B. Chan, P.J. Hall, Scale adjustments for clasifiers in high-dimensional, low sample size settings, Biometrika 96 (2009) 469–478.
[12] L. Clemmensen, T. Hastie, D. Witten, B. Ersbøll, Sparse discriminant analysis, Technometrics 53 (2011) 406–413.
[13] G.B. Cybis, M. Valk, S.R.C. Lopes, Clustering and classification problems in genetics through U-statistics, J. Stat. Comput. Simul. 88 (10) (2018) 1882–1902.
[14] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discriminant methods for the classification of tumors using gene expression data, J. Amer. Statist. Assoc. 97 (2002) 77–87.
[15] J. Fan, Y. Fan, High-dimensional classification using features annealed independence rules, Ann. Statist. 36 (2008) 2605–2637.
[16] Y. Fan, Y. Kong, D. Li, Z. Zheng, Innovated interaction screening for high-dimensional nonlinear classification, Ann. Statist. 43 (2015) 1243–1272.
[17] P.J. Hall, Y. Pittelkow, M. Ghosh, Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008) 159–173.
[18] N. Hao, B. Dong, J. Fan, Sparsifying the Fisher linear discriminant by rotation, J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (2015) 827–851.
[19] S. Huang, T. Tong, H. Zhao, Bias-corrected diagonal discriminant rules for high-dimensional classification, Biometrics 66 (2010) 1096–1106.
[20] M. Hyodo, T. Yamada, T. Himeno, T. Seo, A modified linear discriminant analysis for high-dimensional data, Hiroshima Math. J. 42 (2012) 209–231.
[21] B. Jiang, X. Wang, C. Leng, A direct approach for sparse quadratic discriminant analysis, arXiv:1510.00084v3, 2016.
[22] T.Y. Kim, Z.-M. Luo, C. Kim, The central limit theorem for degenerate variable U-statistics under dependence, J. Nonparametr. Stat. 23 (2011) 683–699.
[23] V.S. Koroljuk, Y.V. Borovskich, Theory of U-Statistics, Kluwer Academic Press, Dordrecht, 1994.
[24] A.J. Lee, U-Statistics: Theory and Practice, CRC Press, Boca Raton, FL, 1990.
[25] E.L. Lehmann, Elements of Large-Sample Theory, Springer, New York, 1999.
[26] Q. Li, J. Shao, Sparse quadratic discriminant analysis for high-dimensional data, Statist. Sinica 25 (2015) 457–473.
[27] R. Luo, X. Qi, Asymptotic optimality of sparse linear discriminant analysis with arbitrary number of classes, Scand. J. Stat. 44 (2017) 598–616.
[28] Q. Mai, A review of discriminant analysis in high dimensions, WIREs Comput. Statist. 5 (2013) 190–197.
[29] T. Mikosch, Weak invariance principle for weighted U-statistics with varying kernels, J. Multivariate Anal. 47 (1993) 82–102.
[30] A. Pinheiro, P.K. Sen, H. Pinheiro, Decomposibility of high-dimensional diversity measures: Quasi U-statistics, martigales, and nonstandard asymptotics, J. Multivariate Anal. 100 (2009) 1645–1656.
[31] S.R. Searle, Linear Models, Wiley, New York, 1971.
[32] G.A.F. Seber, Multivariate Observations, Wiley, New York, 2004.
[33] R.J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, Weinheim, 1980.
[34] J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis by thresholding for high dimensional data, Ann. Statist. 39 (2011) 1241–1265.

[35] M.A. Shipp, K.N. Ross, P. Tamayo, et al., Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68–74.
[36] A. Statnikov, I. Tsamardinos, Y. Dosbayev, C.F. Aliferis, GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data, Int. J. Med. Inform. 74 (2005) 491–503.
[37] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[38] C. Wang, B. Jiang, On the dimension effect of regularized linear discriminant analysis, arXiv:1710.03136v1, 2017.
[39] D. Witten, R.J. Tibshirani, Penalized classification using Fisher's linear discrimininant, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011) 753–772.
[40] P.-S. Zhong, S.X. Chen, Tests for high-dimensional regression coefficients with factorial designs, J. Amer. Statist. Assoc. 106 (2011) 260–274.