On Some Fast And Robust Classifiers For High Dimension, Low Sample Size Data

Sarbojit Roy
Indian Institute of Technology
Kanpur, India
sarbojit@iitk.ac.in

Jyotishka Ray Choudhury
Indian Statistical Institute
Kolkata, India
bs1903@isical.ac.in

Subhajit Dutta
Indian Institute of Technology
Kanpur, India
duttas@iitk.ac.in

Abstract

In high dimension, low sample size (HDLSS) settings, *distance concentration* phenomena affects the performance of several popular classifiers which are based on Euclidean distances. The behaviour of these classifiers in high dimensions is completely governed by the first and second order moments of the underlying class distributions. Moreover, the classifiers become absolutely useless for such HDLSS data when the first two moments of the competing distributions are equal, or when the moments do not exist. In this work, we propose robust, computationally efficient and tuning-free classifiers applicable in HDLSS scenarios. As the data dimension increases, these classifiers yield *perfect classification* if the one-dimensional marginals of the underlying distributions are different. We also establish strong theoretical properties for the proposed classifiers in *ultrahigh-dimensional* settings. Numerical experiments with a wide variety of simulated examples and analysis of real data sets exhibit clear and convincing advantages over existing methods.

1 INTRODUCTION

Let us consider a classification problem involving two distribution functions $\mathbf{F}_1$ and $\mathbf{F}_2$ on $\mathbb{R}^p$ with $p \geq 1$. Suppose $\chi_1 = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ and $\chi_2 = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$ are random samples drawn from $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively and $\chi = \chi_1 \cup \chi_2$ is the training sample of size $n = n_1 + n_2$. The prior probability of j-th class is given by $\pi_j (> 0)$ for $j = 1, 2$ with $\pi_1 + \pi_2 = 1$. Using the training sample, a classifier δ assigns a new point $\mathbf{z} \in \mathbb{R}^p$ to one of the two competing classes. We develop classifiers that yield *perfect classification* under fairly general conditions in high dimension, low sample size (HDLSS) settings, where the sample size n remains fixed, but the dimension p increases. The classifier δ is said to yield *perfect classification* in HDLSS settings if the misclassification probability of δ goes to 0 as $p \to \infty$.

> In the classical setting, p is fixed and $n \to \infty$. Information is accumulated as more samples are collected.
>
> In the HDLSS settings, n is fixed while $p \to \infty$. Information is accumulated as more features are measured.

1.1 Literature Review

In the HDLSS asymptotic regime, Euclidean distance (ED) based classifiers face some natural drawbacks due to *distance concentration* (Aggarwal et al., 2001; Francois et al., 2007). To give a mathematical exposition of this fact, let $\boldsymbol{\mu}_j$ and Σ_j denote the mean vector and the covariance matrix of $\mathbf{F}_j$ for $j = 1, 2$. We assume that the following limits exist:

$$\nu^2 = \lim_{p \to \infty} \frac{1}{p} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \text{ and}$$
$$\sigma_j^2 = \lim_{p \to \infty} \frac{1}{p} tr(\Sigma_j) \text{ for } j = 1, 2. \qquad (1.1)$$

Here, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^p$ and $tr(M)$ denotes the trace of a matrix M. The constants ν^2 and $|\sigma_1^2 - \sigma_2^2|$ can be interpreted as asymp-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

totic measures of the difference between locations and scales of $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively. Hall et al. (2005) studied the consequence of distance concentration on some popular ED based classifiers such as the 1-nearest neighbor (1NN) classifier (Hastie et al., 2009), average distance (AVG) classifier (Chan and Hall, 2009b) and support vector machines (SVM) (Vapnik, 1998). The authors showed that in high dimensions, these methods are incapable of correctly classifying an observation if the location difference between the competing populations gets masked by their difference in scales, i.e., $\nu^2 < |\sigma_1^2 - \sigma_2^2|$. Chan and Hall (2009b); Dutta and Ghosh (2016) proposed some improved classifiers that yield *perfect classification* if $\nu^2 > 0$, or $\sigma_1^2 \neq \sigma_2^2$. However, these improved methods fail in high dimensions when the competing populations have same location and scale, i.e., $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, or when ν^2, σ_1^2 and σ_2^2 do not exist. The limitations of these methods stem from the fact that they are based on Euclidean distances, and the behavior of ED in the HDLSS asymptotic regime is completely governed by these constants. As a result, ED based classifiers cannot distinguish between populations that do not have differences in their first two moments. On top of that, these classifiers lack robustness since ED is sensitive to outliers. Chan and Hall (2009a) proposed a robust version of the NN classifier for high-dimensional data, but it is only applicable to a specific type of two-class location problem. Other approaches for classifying high-dimensional data include Globerson and Roweis (2005); Tomašev et al. (2014); Weinberger and Saul (2009). A recent work by Thrampoulidis (2020) discusses the high-dimensional behavior of several classifiers, but under an additional assumption of Gaussianity.

1.2 Motivation

Li and Zhang (2020) proposed a method for testing equality of two distributions, where the authors considered a new measure of distance between $\mathbf{F}_1$ and $\mathbf{F}_2$ as defined below:

$$\tau = \mathrm{E}\big[h(\mathbf{X}_1, \mathbf{X}_2) + h(\mathbf{Y}_1, \mathbf{Y}_2) - 2h(\mathbf{X}_1, \mathbf{Y}_1)\big].$$

Here, $h : \mathbb{R}^p \times \mathbb{R}^p \to [-1, 1]$ is given by

$$h(\mathbf{u}, \mathbf{v}) = \frac{1}{2\pi} \sin^{-1}\left(\frac{1 + \mathbf{u}^\top \mathbf{v}}{[(1 + \|\mathbf{u}\|^2)(1 + \|\mathbf{v}\|^2)]^{\frac{1}{2}}}\right)$$

for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ with $p \geq 1$. The authors showed that for a fixed p, $\tau = 0$ iff $\mathbf{F}_1 = \mathbf{F}_2$. This property of τ is useful for distinguishing one distribution from another, and can be utilized in classification problems as well. However, a classifier that directly utilizes τ. faces certain challenges in the HDLSS setting.

To motivate the problem, we modify the scale-adjusted average distance (SAVG) classifier (Chan and Hall, 2009b) by simply replacing the squared Euclidean norm $\|\mathbf{u} - \mathbf{v}\|^2$ with $h(\mathbf{u}, \mathbf{v})$ defined above. A formal definition of this modified classifier (henceforth, referred to as δ_0) is given in Section 2, where we also discuss how δ_0 uses τ to classify a test observation.

Let us now consider the following examples:

Example 1 $X_{1k} \overset{i.i.d.}{\sim} N(1, 1)$ *and* $Y_{1k} \overset{i.i.d.}{\sim} N(1, 2)$,

Example 2 $X_{1k} \overset{i.i.d.}{\sim} N(0, 3)$ *and* $Y_{1k} \overset{i.i.d.}{\sim} t_3$,

for $1 \leq k \leq p$. Here, $N(\mu, \sigma^2)$ denotes the univariate Gaussian distribution with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma(> 0)$, and t_κ denotes the standard Student's t distribution with $\kappa(> 0)$ degrees of freedom. In Figure 1, we compare the performance of the classifier δ_0 with some popular classifiers like 1NN, the usual SAVG, SVM with the linear kernel (SVM-LIN) and SVM with the radial basis function (SVM-RBF) kernel. Details of the simulation study is given in Section 4.

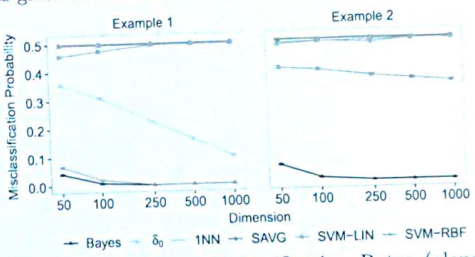

Figure 1: Average Misclassification Rates (along with Standard Errors) of δ_0 and Some Popular Classifiers. Dimension is in Logarithmic Scale.

In the first example, $\nu^2 = 0$ (since $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{1}_p$) but $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$. The classifier δ_0 identifies this difference in scales and yields a moderate performance. Whereas the existing classifiers (except SVM-RBF) misclassify 50% of the observations. SVM-RBF capitalizes on the difference between σ_1^2 and σ_2^2, and perfectly classifies the test observations as dimension increases. The problem of classification is more challenging in **Example 2**. Here, we have $\nu^2 = 0$ (since $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$) and $\sigma_1^2 = \sigma_2^2 = 3$, i.e., there is no difference between either of the location and scale parameters. As a result, δ_0 along with existing classifiers fail to correctly classify the test observations. We will revisit these examples again in Sections 3.1.2 and 4.

1.3 Our Contribution

In this article, we develop classifiers that are suitable for high-dimensional data. The behavior of the proposed classifiers in HDLSS settings do not depend on the existence of the moments. If the one-dimensional marginals of the underlying populations are different, then the proposed classifiers are shown to yield *perfect classification* in HDLSS settings.

The proposed classifiers

- are robust,
- computationally fast,
- free from tuning parameters, and
- have strong theoretical properties.

The rest of the article is organized as follows. In Section 2, we propose a classifier and further modify it to achieve improved classification accuracy under specific conditions. Asymptotic properties of the proposed classifiers are studied in Section 3. A theoretical result is presented in Section 3.1.2 to analyze their relative performances. In Section 3.2, we investigate their behavior when both n and p increase. Performance of the classifiers is studied through a numerical exercise involving several simulated data sets in Section 4. We also examine the behavior of the classifiers on real data sets in Section 5. The article ends with concluding remarks in Section 6. All proofs and relevant mathematical details are provided in Supplementary A. Additional details of the numerical performance of the classifiers and a link to related R codes can be found in Supplementary B.

2 METHODOLOGY

Let us recall the classifier δ_0 mentioned in Section 1.2. Fix $\mathbf{z} \in \mathbb{R}^p$. For given random samples $\mathcal{X}_1$ and $\mathcal{X}_2$ with sizes $n_1 (\geq 2)$ and $n_2 (\geq 2)$, respectively, the classifier δ_0 is formally defined as

$$\delta_0(\mathbf{z}) = \arg\min_{j \in \{1,2\}} L_j(\mathbf{z}), \text{ where } L_j(\mathbf{z}) = T_{jj} - 2T_j(\mathbf{z}),$$

$$T_{jj} = \frac{1}{n_j(n_j-1)} \sum_{\substack{\mathbf{U},\mathbf{U}' \in \mathcal{X}_j, \\ \mathbf{U} \neq \mathbf{U}'}} \sum h(\mathbf{U},\mathbf{U}') \text{ and}$$

$$T_j(\mathbf{z}) = \frac{1}{n_j} \sum_{\mathbf{U} \in \mathcal{X}_j} h(\mathbf{U},\mathbf{z}) \text{ for } j = 1,2. \tag{2.1}$$

For a random vector $\mathbf{Z}$ (independent from $\mathcal{X}$), the misclassification probability of a classifier δ is defined as $\Delta = \mathrm{P}[\delta(\mathbf{Z}) \neq \text{ true class label of } \mathbf{Z}]$. Throughout this article, we will follow this definition of the misclassfication probability. In the previous section, we have introduced the constants ν^2, σ_1^2 and σ_2^2. Now, we define $\nu_{jj'} = \lim_{p\to\infty} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'}/p$ for $j, j' \in \{1,2\}$ and further assume the following:

(i) There exists a constant C_0 such that $E[|U_k|^4] < C_0 < \infty$ for all $1 \leq k \leq p$, where $\mathbf{U} = (U_1, \ldots, U_p)^\top \sim \mathbf{F}_j$ for $j = 1, 2$.

(ii) The constants $\nu_{jj'}$ and σ_j^2 exist for $j, j' \in \{1,2\}$.

Let $\mathbf{U}$ and $\mathbf{V}$ be two independent vectors such that $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$ for $j, j' \in \{1,2\}$. We also assume that the components of the sequence $\{U_k V_k, k \geq 1\}$ are weakly dependent. In particular,

(iii) $\sum\sum_{1 \leq k < k' \leq p} \mathrm{Corr}(U_k V_k, U_{k'} V_{k'}) = o(p^2)$.

Assumption (iii) is trivially satisfied if the component variables of the underlying populations are independent. It continues to hold with some additional conditions on their dependence structure. For example, (iii) is satisfied when the sequence $\{U_k V_k, k \geq 1\}$ has ρ-mixing property (Bradley, 2005; Hall et al., 2005). Conditions similar to (iii) are frequently considered in the literature for studying high-dimensional behavior of various statistical procedures (Aoshima et al., 2018).

Lemma 2.1 *Suppose assumptions* (i)-(iii) *are satisfied. Let* $L(\mathbf{Z}) = L_2(\mathbf{Z}) - L_1(\mathbf{Z})$ *for a test observation* $\mathbf{Z}$.

(a) If $\mathbf{Z} \sim \mathbf{F}_1$, *then* $|L(\mathbf{Z}) - \tau| \xrightarrow{\mathrm{P}} 0$ *as* $p \to \infty$.

(b) If $\mathbf{Z} \sim \mathbf{F}_2$, *then* $|L(\mathbf{Z}) + \tau| \xrightarrow{\mathrm{P}} 0$ *as* $p \to \infty$.

Lemma 2.1 states that if $\mathbf{Z} \sim \mathbf{F}_1$ ($\mathbf{F}_2$, respectively), then the discriminant of δ_0 converges in probability to τ, a positive (negative, respectively) quantity as $p \to \infty$. Let Δ_0 denote the misclassification probability of the classifier δ_0. The following theorem shows that in HDLSS settings, the asymptotic behavior of δ_0 is governed by the constants $\nu_{jj'}$, and σ_j^2 for $j, j' \in \{1,2\}$.

Theorem 2.2 *Suppose that assumptions* (i)-(iii) *are satisfied, and either of the following conditions hold:*

(a) ν_{11}, ν_{12} *and* ν_{22} *are unequal,*

(b) $\nu_{11} = \nu_{12} = \nu_{22} \neq 0$ *and* $\sigma_1^2 \neq \sigma_2^2$.

For any $\pi_1 > 0$, $\Delta_0 \to 0$ *as* $p \to \infty$.

Theorem 2.2 shows that if $\mathbf{F}_1$ and $\mathbf{F}_2$ differ either in their locations and/or scales, then Δ_0 converges to 0 as dimension increases. Recall **Example 1**, and note that condition (b) of Theorem 2.2 is satisfied in this example since $|\sigma_1^2 - \sigma_2^2| = 1$. In **Example 2**, both (a) and (b) are violated and Theorem 2.2 fails to hold. This gives us a clear explanation why the classifier δ_0 performed well in the first example, but failed in the second one (see Figure 1). Now, we develop some classifiers whose asymptotic properties are not governed by the constants $\nu_{jj'}$, and σ_j^2 for $j, j' \in \{1, 2\}$. The proposed classifiers use differences between the one-dimensional marginals of $\mathbf{F}_1$ and $\mathbf{F}_2$, and attain *perfect classification* in high dimensions under fairly general conditions.

2.1 A New Measure of Distance

Let $F_{j,k}$ denote the distribution of the random variable U_k, where $\mathbf{U} = (U_1, \ldots, U_p)^\top \sim \mathbf{F}_j$ for $j = 1, 2$ and $1 \le k \le p$. Suppose, $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} \mathbf{F}_1$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} \mathbf{F}_2$. Fix $1 \le k \le p$ and recall the definition of τ_k in Section 1.2. The distance between $F_{1,k}$ and $F_{2,k}$ is given by $\tau_k = \mathrm{E}[h(X_{1k}, X_{2k}) - 2h(X_{1k}, Y_{1k}) + h(Y_{1k}, Y_{2k})]$. Here, $\tau_k \ge 0$ and equality holds iff $F_{1,k} = F_{2,k}$. We denote the average of these distances by $\bar{\tau}_p = \sum_{k=1}^p \tau_k / p$. Clearly, $\bar{\tau}_p = 0$ iff $\tau_k = 0$ for all $1 \le k \le p$,

$$\text{i.e., } \bar{\tau}_p = 0 \text{ iff } F_{1,k} = F_{2,k} \text{ for all } 1 \le k \le p.$$

This property of $\bar{\tau}_p$ suggests that it can be used as a *measure of separation* between $\mathbf{F}_1$ and $\mathbf{F}_2$. If $F_{1,k} \ne F_{2,k}$ for some $1 \le k \le p$, then $\bar{\tau}_p$ is strictly positive. This is the fundamental idea that we will use in developing a new classifier.

Recall the definition of h given in Section 1.2, and consider

$$\bar{h}_p(\mathbf{u}, \mathbf{v}) = \frac{1}{p} \sum_{k=1}^{p} h(u_k, v_k) \text{ for } \mathbf{u}, \mathbf{v} \in \mathbb{R}^p. \tag{2.2}$$

Using (2.2), we re-write the definition of $\bar{\tau}_p$ as

$$\bar{\tau}_p = \mathrm{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{X}_2) - 2\bar{h}_p(\mathbf{X}_1, \mathbf{Y}_1) + \bar{h}_p(\mathbf{Y}_1, \mathbf{Y}_2)].$$

Let $\bar{\tau}_p(1,1), \bar{\tau}_p(1,2) (= \bar{\tau}_p(2,1))$ and $\bar{\tau}_p(2,2)$ denote the quantities $\mathrm{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{X}_2)]$, $\mathrm{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{Y}_1)]$ and $\mathrm{E}[\bar{h}_p(\mathbf{Y}_1, \mathbf{Y}_2)$, respectively. Observe that

$$\bar{\tau}_p = \bar{\tau}_p(1,1) - 2\bar{\tau}_p(1,2) + \bar{\tau}_p(2,2). \tag{2.3}$$

Fix $\mathbf{z} \in \mathbb{R}^p$. We define the following:

$$\bar{T}_{jj} = \frac{1}{n_j(n_j - 1)} \sum_{\substack{\mathbf{U}, \mathbf{U}' \in \chi_j \\ \mathbf{U} \ne \mathbf{U}'}} \sum \bar{h}_p(\mathbf{U}, \mathbf{U}'),$$

$$\bar{T}_j(\mathbf{z}) = \frac{1}{n_j} \sum_{\mathbf{U} \in \chi_j} \bar{h}_p(\mathbf{U}, \mathbf{z}), \; \bar{L}_j(\mathbf{z}) = \bar{T}_{jj} - 2\bar{T}_j(\mathbf{z})$$

$$\text{for } j = 1, 2 \text{ and } \bar{L}(\mathbf{z}) = \bar{L}_2(\mathbf{z}) - \bar{L}_1(\mathbf{z}). \tag{2.4}$$

It follows from the above definitions that

$$\mathrm{E}[\bar{T}_j(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_{j'}] = \bar{\tau}_p(j, j') \text{ and}$$
$$\mathrm{E}[\bar{T}_{jj}] = \bar{\tau}_p(j, j) \text{ for } j, j' \in \{1, 2\}. \tag{2.5}$$

Consequently, we obtain

$$\mathrm{E}[\bar{L}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_1] = \bar{\tau}_p \ge 0 \text{ and}$$
$$\mathrm{E}[\bar{L}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_2] = -\bar{\tau}_p \le 0. \tag{2.6}$$

It is clear from the above equation that $\mathrm{E}[\bar{L}(\mathbf{Z})]$ is an indicator to whether a test observation $\mathbf{Z}$ belongs to the first or the second class. This observation motivates us to use $\bar{L}(\mathbf{Z})$ as a discriminant.

2.1.1 A Classifier Based on $\bar{\tau}_p$

Based on (2.6), we propose the following classifier:

$$\delta_1(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{L}(\mathbf{z}) > 0, \\ 2, & \text{otherwise.} \end{cases} \tag{2.7}$$

The classifier δ_1 can also be expressed as $\arg\min_{j \in \{1,2\}} \bar{L}_j(\mathbf{z})$. For given random samples $X_1, \ldots, X_J$ (with $J \ge 2$), we propose $\delta_1(\mathbf{z}) = \arg\min_{1 \le j \le J} \bar{L}_j(\mathbf{z})$, where $\bar{L}_j(\mathbf{z}), \bar{T}_j(\mathbf{z})$ and $\bar{T}_{jj}$ are as defined in (2.4) for $1 \le j \le J$. We denote the misclassification probability of δ_1 by Δ_1.

2.2 Limitations of $\bar{\tau}_p$

To classify a test point, the classifier δ_1 leverages on the quantity $\bar{\tau}_p$, the average of distances between $F_{1,k}$ and $F_{2,k}$ for $1 \le k \le p$. However, $\bar{\tau}_p$ has some limitations. Recall that

$$\begin{aligned} \bar{\tau}_p &= \bar{\tau}_p(1,1) - 2\bar{\tau}_p(1,2) + \bar{\tau}_p(2,2) \\ &= \{\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)\} + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)\}. \end{aligned}$$

Since $\bar{\tau}_p \ge 0$, we always have $\bar{\tau}_p(1,2) \le \{\bar{\tau}_p(1,1) + \bar{\tau}_p(2,2)\}/2$. Without loss of generality, let us assume that $\bar{\tau}_p(1,1) < \bar{\tau}_p(2,2)$. If $\bar{\tau}_p(1,2)$ lies between $\bar{\tau}_p(1,1)$ and $\bar{\tau}_p(2,2)$, i.e., $\bar{\tau}_p(1,1) < \bar{\tau}_p(1,2) < \bar{\tau}_p(2,2)$, then $\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2) < 0$ and $\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2) > 0$. Adding them up may cancel each other. As a result, $\bar{\tau}_p$ may become small despite $\mathbf{F}_1$ and $\mathbf{F}_2$ being different. One way to rectify this problem is to square the two quantities before adding them up. This eliminates the possibility of cancellations. Define

$$\bar{\psi}_p = \{\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)\}^2 + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)\}^2.$$

It is easy to check that $\bar{\psi}_p = 0$ iff $F_{1,k} = F_{2,k}$ for all $1 \le k \le p$. Hence, $\bar{\psi}_p$ can be viewed as a measure of separation between $\mathbf{F}_1$ and $\mathbf{F}_2$. This new measure can also be expressed as

$$\bar{\psi}_p = \frac{1}{2}\left[\bar{\tau}_p^2 + \{\bar{\tau}_p(1,1) - \bar{\tau}_p(2,2)\}^2\right]. \tag{2.8}$$

Observe that if $\bar{\tau}_p(1,2)$ lies between $\bar{\tau}_p(1,1)$ and $\bar{\tau}_p(2,2)$, then $|\bar{\tau}_p(1,1) - \bar{\tau}_p(2,2)| > \bar{\tau}_p$. As a result,

$$\bar{\psi}_p = \frac{1}{2}\left[\bar{\tau}_p^2 + \{\bar{\tau}_p(1,1) - \bar{\tau}_p(2,2)\}^2\right] > \frac{1}{2}\left[\bar{\tau}_p^2 + \bar{\tau}_p^2\right] = \bar{\tau}_p^2.$$

Clearly, $\bar{\tau}_p \ge 1$ implies $\bar{\psi}_p > \bar{\tau}_p$, making $\bar{\psi}_p$ a better choice than $\bar{\tau}_p$ (in terms of measuring separation between two distributions). In general, if the underlying distributions $\mathbf{F}_1$ and $\mathbf{F}_2$ are such that $\bar{\tau}_p(1,2) > \min\{\bar{\tau}_p(1,1), \bar{\tau}_p(2,2)\}$, then a classifier that utilizes $\bar{\psi}_p$ is shown to have better classification accuracy than the classifier δ_1 (see Section 3.1.2 for more details). The modification proposed in (2.8) is similar to what Biswas and Ghosh (2014) had suggested for improving the power of some energy based tests for HDLSS data.

2.2.1 A Classifier Based on $\bar{\psi}_p$

We now develop a classifier that leverages the amplified measure of dissimilarity $\bar{\psi}_p$. First, let us estimate $\bar{\tau}_p(1,2)$ by

$$\bar{T}_{12} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \bar{h}_p(\mathbf{X}_i, \mathbf{Y}_j). \tag{2.9}$$

Fix $\mathbf{z} \in \mathbb{R}^p$. We define

$$\bar{\theta}(\mathbf{z}) = \frac{1}{2}\{\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}\}\{\bar{L}_2(\mathbf{z}) - \bar{L}_1(\mathbf{z})\} + \frac{1}{2}\{\bar{T}_{22} - \bar{T}_{11}\}\{\bar{L}_2(\mathbf{z}) + \bar{L}_1(\mathbf{z}) + 2\bar{T}_{12}\}. \tag{2.10}$$

We will prove that $|\bar{\theta}(\mathbf{Z})|$ is a consistent estimator of $\bar{\psi}_p$ where $\mathbf{Z}$ is a test observation. In particular, $\bar{\theta}(\mathbf{Z})$ converges in probability to $\bar{\psi}_p$ as $p \to \infty$ if $\mathbf{Z} \sim \mathbf{F}_1$, and to $-\bar{\psi}_p$ if $\mathbf{Z} \sim \mathbf{F}_2$ (see Section 3 for more details). This motivates us to propose the following classifier:

$$\delta_2(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{\theta}(\mathbf{z}) > 0, \\ 2, & \text{otherwise.} \end{cases} \tag{2.11}$$

Let Δ_2 denote the misclassfication probability of the classifier δ_2. Unlike δ_1, the classifier δ_2 cannot be readily extended to deal with J class problems when $J > 2$. For multi-class problems, we implement the idea of 'majority voting' (Friedman et al., 2001).

Examples 1 and **2** establish the necessity of the modified measure $\bar{\psi}_p$ and the advantage of using δ_2 over δ_1. In Figure 2, we see that δ_2 has substantial improvement over δ_1 in terms of misclassfication probability. This improvement is due to the fact that

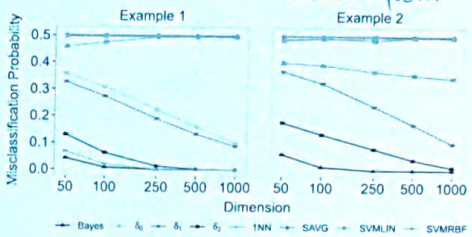

Figure 2: Average Misclassification Rates (along with Standard Errors) of the Proposed Classifiers.

$\bar{T}_{12}$ lies between $\bar{T}_{11}$ and $\bar{T}_{22}$ in both examples (see Table 2 in the Supplementary B). A theoretical result on the relative performance of the classifiers δ_1 and δ_2 is presented in Section 3.1.2.

3 ASYMPTOTIC PROPERTIES

In HDLSS settings, n is fixed and $p \to \infty$, whereas in the *ultrahigh-dimensional* setting, p grows simultaneously with n. The behavior of the classifiers δ_1 and δ_2 is investigated in both aymptotic regimes. We first show that the classifiers yield *perfect classification* in HDLSS settings under fairly general conditions.

3.1 Asymptotic Behavior in HDLSS Settings

Suppose $\mathbf{U}$ and $\mathbf{V}$ are two independent vectors such that $\mathbf{U} = (U_1, \ldots, U_p)^\top \sim \mathbf{F}_j$ and $\mathbf{V} = (V_1, \ldots, V_p)^\top \sim \mathbf{F}_{j'}$ for $j, j' \in \{1, 2\}$. We assume that the component variables are weakly dependent. In particular, we assume

A1. $\sum\sum_{1 \le k < k' \le p} \mathrm{Corr}(h(U_k, V_k), h(U_{k'}, V_{k'})) = o(p^2)$,

where h is defined in Section 1.2. Assumption A1 is trivially satisfied if the component variables of the underlying distributions are independently distributed and it continues to hold when the components have weak dependence among them. For example, A1 is satisfied when the sequence $\{h(U_k, V_k), k \ge 1\}$ has ρ-mixing property. Note that if the sequences $\{U_k, k \ge 1\}$ and $\{V_k, k \ge 1\}$

have ρ-mixing property, then $\{h(U_k, V_k), k \geq 1\}$ has ρ-mixing property for every measurable function h (see Theorem 6.6-II of Bradley (2007)).

Recall assumption (iii) introduced in Section 2. Both (iii) and A1 require the component variables to be weakly dependent. However, A1 is weaker between the two since, unlike (iii), it does not require the existence of first and second order moments of the distributions. The boundedness of h enables A1 to deal with heavy-tailed distribution.

Lemma 3.1 *If* A1 *is satisfied, then for a test observation* $\mathbf{Z}$*, we have*

(a) If $\mathbf{Z} \sim \mathbf{F}_1$*, then* $|\bar{L}(\mathbf{Z}) - \bar{\tau}_p| \xrightarrow{\text{P}} 0$ *and* $|\bar{\theta}(\mathbf{Z}) - \bar{\psi}_p| \xrightarrow{\text{P}} 0$ *as* $p \to \infty$.

(b) If $\mathbf{Z} \sim \mathbf{F}_2$*, then* $|\bar{L}(\mathbf{Z}) + \bar{\tau}_p| \xrightarrow{\text{P}} 0$ *and* $|\bar{\theta}(\mathbf{Z}) + \bar{\psi}_p| \xrightarrow{\text{P}} 0$ *as* $p \to \infty$.

Similar results on distance concentration can be derived for independently distributed sub-Gaussian components (see Theorem 3.1.1 of Vershynin (2018) for further details). Lemma 3.1 is stronger than existing results in the sense that it holds even when the components are not necessarily independent, or sub-Gaussian.

Lemma 3.1 states that both the discriminants converge in probability to a non-negative value if $\mathbf{Z} \sim \mathbf{F}_1$, while they converge in probability to a value which is not positive when $\mathbf{Z} \sim \mathbf{F}_2$. Now, we expect δ_1 and δ_2 to have a good performance if $\bar{\tau}_p$ and $\bar{\psi}_p$ do not vanish with increasing dimension. Clearly, $\bar{\tau}_p = \bar{\psi}_p = 0$ iff $F_{1,k} = F_{2,k}$ for all $1 \leq k \leq p$. Hence, it is reasonable to assume the following:

A2. $\liminf_p \bar{\tau}_p > 0$.

A2 implies that the separation between $\mathbf{F}_1$ and $\mathbf{F}_2$ is asymptotically non-negligible. Observe that this assumption is satisfied if the component variables of $\mathbf{U} \sim \mathbf{F}_j$ are identically distributed for $j = 1, 2$. Then, $\tau_k = \tau_1 > 0$ for all $k \geq 1$, making $\bar{\tau}_p (= \tau_1)$ free of p. It follows from the definition of $\bar{\psi}_p$ in (2.8) that A2 implies $\liminf_p \bar{\psi}_p > 0$.

3.1.1 Asymptotic Properties of δ_1 and δ_2 in HDLSS Settings

Now, we discuss behavior of the classifiers δ_1 and δ_2 in HDLSS settings. We show that under fairly general conditions, the proposed classifiers δ_1 and δ_2 perfectly classify an observation as the dimension increases.

Theorem 3.2 *If* A1 *and* A2 *are satisfied, then for any* $\pi_1 > 0$*,*

(a) $\Delta_1 \to 0$ *as* $p \to \infty$*, and*

(b) $\Delta_2 \to 0$ *as* $p \to \infty$.

Observe that the asymptotic behavior of the classifiers are no longer governed by the constants $\nu_{jj'}$, σ_j^2 for $j, j' \in \{1, 2\}$. In fact, their behavior do not depend on the existence of moments. In this sense, the classifiers δ_1 and δ_2 are robust.

> Asymptotic behavior of the proposed classifiers is free of moment conditions.
>
> The classifiers yield *perfect classificaton* under weak conditions.

One should observe that assumptions A1 and A2 are fairly general, and Theorem 3.2 is stronger than what exists in the current literature.

3.1.2 Comparison Between δ_1 and δ_2

We have seen that both the proposed classifiers yield *perfect classification* under the same set of assumptions. The next result provides a set of sufficient conditions under which one classifier performs better than the other.

First, let us consider the following assumption:

A3. There exists a $p_0 \in \mathbb{N}$ such that $\bar{\tau}_p(1,2) > \min\{\bar{\tau}_p(1,1), \bar{\tau}_p(2,2)\}$ for all $p \geq p_0$.

If assumption A3 is satisfied, then one of $\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)$ and $\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)$ is positive, while the other is negative. $\bar{\tau}_p$ may take a small value (see the discussion in Section 2.2). The next result suggests that under such circumstances, δ_2 leads to better performance than δ_1.

Theorem 3.3 *If assumptions* (A1) – (A3) *are satisfied, then there exists an integer* p_0' *such that*

$$\Delta_2 \leq \Delta_1 \text{ for all } p \geq p_0'.$$

Recall that $\bar{T}_{11}, \bar{T}_{12}$ and $\bar{T}_{22}$ are unbiased estimators of $\bar{\tau}_p(1,1), \bar{\tau}_p(1,2)$ and $\bar{\tau}_p(2,2)$, respectively (see (2.5)). We use these estimators to understand the

relative performance of the classifiers. In **Examples 1** and **2**, $\bar{T}_{12}$ is observed to be lying between $\bar{T}_{11}$ and $\bar{T}_{22}$ (see Table 2 in Supplementary B). Following Theorem 3.3, we expect Δ_2 to be smaller than Δ_1 in these examples. Figure 2 shows that the estimated misclassfication probability of the classifier δ_2 is indeed lower than that of δ_1 in both the examples.

3.2 Asymptotic Properties of δ_1 and δ_2 when Sample Size Increases

In this section, we assess the performance of our classifiers in the *ultrahigh-dimensional* asymptotic regime, when the dimension p ($\equiv p_n$) is allowed to grow with n (in non-polynomial order). In particular, we assume the following:

A4. There exists a $0 \leq \beta < 1$ such that

$$\log p_n = O(n^{\beta}).$$

Recall that in the classical asymptotic regime, p is fixed and $n \to \infty$. Therefore, the classical setting is a special case of the *ultrahigh-dimensional* regime with $\beta = 0$. We also assume that $\lim_{n\to\infty} n_1/n = \pi_1 > 0$.

We first present the 'oracle' versions of our classifiers δ_1 and δ_2. Fix $\mathbf{z} \in \mathbb{R}^p$. If $\mathbf{F}_1$ and $\mathbf{F}_2$ are known, then the 'oracle' version of δ_1 is defined as follows:

$$\delta_1^0(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{L}^0(\mathbf{z}) > 0, \\ 2, & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\bar{L}^0(\mathbf{z}) = \bar{L}_2^0(\mathbf{z}) - \bar{L}_1^0(\mathbf{z})$, with $\bar{L}_j^0(\mathbf{z}) = \bar{\tau}_p(j,j) - 2E[\bar{h}_p(\mathbf{U},\mathbf{z})]$ for $\mathbf{U} \sim \mathbf{F}_j$ and $j = 1,2$. Similarly, we define δ_2^0, the 'oracle' version of δ_2 as follows:

$$\delta_2^0(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{\theta}^0(\mathbf{z}) > 0, \\ 2, & \text{otherwise,} \end{cases} \tag{3.2}$$

where $2\bar{\theta}^0(\mathbf{z}) = \bar{\tau}_p \bar{L}^0(\mathbf{z}) + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,1)\} \times \{\bar{L}_2^0(\mathbf{z}) + \bar{L}_1^0(\mathbf{z}) + 2\bar{\tau}_p(1,2)\}$. Note that $\bar{L}(\mathbf{z})$ and $\bar{\theta}(\mathbf{z})$ (defined in (2.4) and (2.10)) are in fact estimators of $\bar{L}^0(\mathbf{z})$ and $\bar{\theta}^0(\mathbf{z})$, respectively.

Let Δ_j^0 denote the misclassification probability of the classifier δ_j^0, $j = 1,2$. In this section, we derive an upper bound on the difference $\Delta_j - \Delta_j^0$ for $j = 1,2$. Furthermore, we show that in the classical setting (i.e., p is fixed), if the competing distributions are absolutely continuous, then $\Delta_j - \Delta_j^0$ converges to 0 for $j = 1,2$ as $n \to \infty$. We first look into the convergence of the discriminants $\bar{L}(\mathbf{z})$ and $\bar{\theta}(\mathbf{z})$.

Lemma 3.4 *Suppose assumption* A4 *is satisfied for some $0 \leq \beta < 1$. For any $\pi_1 > 0$ and $0 < \gamma < (1-\beta)/2$, there exist positive constants B_0 and B_1 such that*

(a) $\mathrm{P}\big[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}\big] \leq O\left(e^{-B_0\{n^{1-2\gamma} - n^{\beta}\}}\right)$,

(b) $\mathrm{P}\big[|\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})| > n^{-\gamma}\big] \leq O\left(e^{-B_1\{n^{1-2\gamma} - n^{\beta}\}}\right)$

for all $\mathbf{z} \in \mathbb{R}^p$.

Since $1 - 2\gamma > \beta$, we have $e^{-\{n^{1-2\gamma} - n^{\beta}\}} \to 0$ as $n \to \infty$. The above result shows that $|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})|$ and $|\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})|$ converge to 0 at an exponential rate as n increases. As a consequence of Lemma 3.4, we have the next result.

Theorem 3.5 *Suppose assumption* A4 *is satisfied for some $0 \leq \beta < 1$. For any $\pi_1 > 0$ and $0 < \gamma < (1-\beta)/2$, there exist positive constants B_0 and B_1 such that*

(a) $\Delta_1 - \Delta_1^0 \leq O\left(e^{-B_0\{n^{1-2\gamma} - n^{\beta}\}}\right) + P\big[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}\big]$,

(b) $\Delta_2 - \Delta_2^0 \leq O\left(e^{-B_1\{n^{1-2\gamma} - n^{\beta}\}}\right) + P\big[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}\big]$.

Clearly, $e^{-B_0\{n^{1-2\gamma} - n^{\beta}\}}$ and $e^{-B_1\{n^{1-2\gamma} - n^{\beta}\}}$ converge to 0 as $n \to \infty$ for all $0 < \gamma < (1-\beta)/2$. Additionally, if $\mathrm{P}[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}]$ and $\mathrm{P}[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}]$ decrease to 0, then Theorem 3.5 suggests that $\Delta_j - \Delta_j^0 \to 0$ as $n \to \infty$ for $j = 1,2$. Observe that in the classical setting, where p is fixed (i.e., $\beta = 0$), if $\mathbf{F}_1, \mathbf{F}_2$ are absolutely continuous, then $\mathrm{P}[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}]$ and $\mathrm{P}[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}]$ go to 0 as $n \to \infty$. Suppose, A4 is satisfied for $\beta > 0$, i.e., p grows with n. One can prove that if assumptions A1 and A2 are satisfied, then $\mathrm{P}[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}]$, $\mathrm{P}[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}]$ go to 0 as $\min\{n, p_n\} \to \infty$. Moreover, Δ_1^0 and Δ_2^0 decay to 0 under the same conditions. As a result, $\Delta_j \to 0$ as $\min\{n, p_n\} \to \infty$ for $j = 1,2$. We do not provide with mathemtical details of the arguments as it is similar to that of Theorem 3.2.

3.3 Computational Complexity

Computing $\bar{T}_{jj'}$ and $\bar{T}_j(\mathbf{z})$ for $\mathbf{z} \in \mathbb{R}^p$ requires $O(n^2 p)$ and $O(np)$ operations, respectively, for $j, j' \in \{1,2\}$. Thus, the overall complexity of classifying an observation using δ_1 and δ_2 is $O(n^2 p)$. Clearly, the complexity increases linearly with respect to p. This makes the methods advantageous in analyzing high-dimensional data sets. We report average time taken by these classifiers to classify a

test observation in Table 2 of the Supplementary B. It clearly shows the advantage of using δ_1 and δ_2 over some popular classifiers.

4 SIMULATION STUDY

In this section, we analyze some simulated data sets to compare the classifiers δ_0, δ_1 and δ_2 with some popular classifiers like GLMNET (Hastie et al., 2009), the usual 1NN, NN based on the random projection method (NN-RAND) (Deegalla and Bostrom, 2006), neural networks (NNET)(Bishop, 1995), SVM-LIN and SVM-RBF. All numerical exercises are performed on an Intel Xeon Gold 6140 CPU (2.30GHz, 2295 Mhz) using the statistical software R. Details about the used packages and parameters related to implementation of the popular classifiers are provided in Supplementary B.

Recall **Examples 1** and **2** introduced in Section 1. Three more examples are considered to conduct a comprarative study of the performance of these classifiers.

Example 3 $X_{1k} \overset{i.i.d.}{\sim} C(0,1)$ *and* $Y_{1k} \overset{i.i.d.}{\sim} C(1,1)$,

Example 4 $X_{1k} \overset{i.i.d.}{\sim} C(0,1)$ *and* $Y_{1k} \overset{i.i.d.}{\sim} C(0,2)$,

Example 5 $X_{1k} \overset{i.i.d.}{\sim} \text{Par}(1,1)$ *and* $Y_{1k} \overset{i.i.d.}{\sim} \text{Par}(1.25,1)$,

for $1 \le k \le p$. Here, $C(\mu,\sigma)$ denotes the Cauchy distribution with location $\mu \in \mathbb{R}$ and scale $\sigma > 0$, while $\text{Par}(\theta, s)$ denotes the Pareto distribution with $\theta > 0$ and scale $s > 0$.

Examples 3, **4** and **5** correspond to location, scale and location-scale problem, respectively. All three examples involve heavy-tailed distributions. In each example, we simulated data for $p = 50$, 100, 250, 500 and 1000. The training sample was formed with 20 observations from each class and a test set of size 200 (100 from each class) was used. This process was repeated 100 times to estimate the misclassification probabilities, which are reported in Figure 3 along with the standard errors.

The performance of δ_0 in **Examples 1** and **2** is already presented in Section 2. Figure 3 shows that δ_0 fails miserably in **Examples 3-5**. Recall Theorem 2.2, and observe that assumption (iii) is violated for these examples since the competing distributions are heavy-tailed. Hence, the poor performance of δ_0.

The classifiers δ_1 and δ_2 lead to promising results in all examples. Assumption A1 is satisfied in these examples since the component variables are independently distributed. Also, the marginals are identical, i.e., $F_{1,k} = F_{1,1}$ and $F_{2,k} = F_{2,1}$ for all $1 \le k \le p$. Therefore, $\bar{\tau}_p (= \tau_1) > 0$ is free of both n and p. Hence, A2 is satisfied. Consequently, Theorem 3.2 holds for all examples.

Figure 3 shows that the misclassification probability of δ_2 is smaller than that of δ_1 in **Examples 1**, **2**, **4** and **5**. Whereas, δ_1 outperformed δ_2 in **Example 3**. Recall the discussion on relative performance of the proposed classifiers in Section 3.1.2. Following Theorem 3.3, we estimated $\bar{\tau}_p(1,1)$, $\bar{\tau}_p(1,2)$, and $\bar{\tau}_p(2,2)$ by $\bar{T}_{11}$, $\bar{T}_{12}$, and $\bar{T}_{22}$, respectively, for all examples and found that $\bar{T}_{12} < \min\{\bar{T}_{11}, \bar{T}_{22}\}$ in **Example 3**, while $\bar{T}_{12} > \min\{\bar{T}_{11}, \bar{T}_{22}\}$ in other examples (see Table 2 of Supplementary B). The numerical evidence clearly support our claim in Theorem 3.3.

In general, all the popular classifiers exhibit poor performance in general, except for a few instances. In **Example 1**, only SVM-RBF identified the difference between scales of the competing populations and yielded *perfect classification*. The rest of the methods failed miserably and misclassified nearly

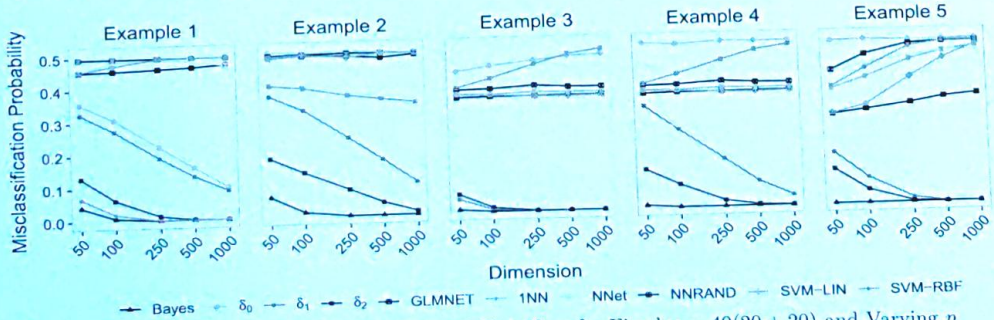

Figure 3: Average Misclassification Rates of Classifiers for Fixed $n = 40(20 + 20)$ and Varying p.

50% of the test observations. In **Example 2**, none of the classifiers had satisfactory results since they are unable to look beyond differences between first two moments in HDLSS settings. In **Examples 3-5**, the competing distributions are heavy-tailed and we observe deteriorating performances of all those popular classifiers.

5 REAL DATA ANALYSIS

We study the performance of the proposed classifiers in two real data sets, namely, `Computers` and `SmoothSubspace` available at the UCR Time Series Archive (see Dau et al., 2018). These data sets have *fixed training* and *test sets*. For our analysis, we combined the training and test data. We randomly selected 50% of the observations from the combined set to form a new set of training observations, while keeping the proportions of observations from different classes consistent. The remaining observations were considered as the test set. This procedure was repeated 100 times to obtain stable estimates of the misclassification probabilities.

The `Computers` (`Comp`) data contains readings on electricity consumption from 251 households in UK, sampled in two-minute intervals over a month. Each observation is of length 720 making the data high-dimensional. Classes are 'Desktop' and 'Laptop' with 250 (125 training and 125 test) samples in each. From Table 1, we observe that δ_0 performed quite poorly, misclassifying almost half of the test observations. The mislassification probability of δ_2 is smaller than that of δ_1 for this data. To understand the relative performance of the classifiers δ_1 and δ_2, we obtained $\bar{T}_{11} = 0.972$, $\bar{T}_{12} = 1.043$, and $\bar{T}_{22} = 1.155$. Observe that $\bar{T}_{12}$ lies between $\bar{T}_{11}$ and $\bar{T}_{22}$. This explains the superior performance of δ_2 over δ_1. In fact, δ_2 outperforms the rest of the classifiers. The regularized linear classifier GLMNET secured the third position with a competitive performance, closely followed by SVM-RBF, whereas 1NN, NNRAND, NNET and SVM-LIN all four classifiers misclassify more than 40% of the observations.

The second data set `SmoothSubspace` (`SSub`) is about testing the ability of a clustering algorithm to extract smooth subspaces for clustering time series data. This data set has 3 classes with 100 (50 train and 50 test) observations each. The observations have dimension 15. We observe in Table 1 that the classifier δ_0 misclassified 18% of the test observations. It is also the worst among all, δ_1 yielded the lowest misclassfication rate, while δ_2 had the second best performance in this data set. We obtained $\bar{T}_{11} = 1.384$, $\bar{T}_{22} = 1.378$, $\bar{T}_{33} = 1.386$, $\bar{T}_{12} = 1.340$, $\bar{T}_{13} = 1.326$, and $\bar{T}_{23} = 1.314$. Observe that $\bar{T}_{jj'} < \min\{\bar{T}_{jj}, \bar{T}_{j'j'}\}$, making both $\bar{T}_{jj} - \bar{T}_{jj'}$ and $\bar{T}_{j'j'} - \bar{T}_{jj'}$ positive for all $1 \leq j \neq j' \leq 3$. Thus, squaring the differences to amplify the measure of separation between $\mathbf{F}_j$ and $\mathbf{F}_{j'}$ becomes unneccessary for all $j \neq j'$. Among the existing methods, NNET produced the highest misclassfication probability. The linear classifiers GLMNET and SVM-LIN also performed poorly, while non-linear classifiers like 1NN, NNRAND and SVM-RBF yielded better misclassfication rates. In particular, SVM-RBF yielded the lowest misclassification rate among the popular classifiers, followed by NN-RAND. Although, their performance is six times worse than that of δ_1.

Table 1: Average Misclassification Rates of Classifiers (in %) with Standard Errors in Parentheses

Data	δ_0	δ_1	δ_2	GLM NET	1NN	NN RAND	NNet	SVM LIN	SVM RBF
`Comp`	47.09	36.40	**35.47**	39.10	42.67	42.04	46.80	46.16	39.95
$J = 2$	(0.24)	(0.22)	(0.21)	(0.24)	(0.28)	(0.27)	(0.28)	(0.34)	(0.27)
`SSub`	18.15	**1.05**	1.33	13.35	8.71	7.09	16.19	10.79	6.35
$J = 3$	(0.27)	(0.06)	(0.08)	(0.28)	(0.20)	(0.22)	(0.44)	(0.28)	(0.19)

6 CONCLUDING REMARKS

In this article, we present classifiers that are capable of perfectly classifying an observation if the one-dimensional marginals of the underlying distributions are different. We study the theoretical properties of the classifiers in the HDLSS and *ultrahigh-dimensional* settings. The proposed classifiers are robust in nature. They yield *perfect classification* even when the competing distributions are heavy-tailed. Their tuning free discriminants make the methods fast and easy to implement. Analysis of several simulated and real data sets show the promising performance of the classifiers.

Suppose the underlying distributions are such that their one-dimensional marginals are identical, and the discriminatory information comes from the joint distributions of groups of variables. Under such circumstances, the proposed measures of separation become 0 and the classifiers may fail to yield *perfect classification*. Developing fast and robust classifiers that harness the distance between joint distributions of components could be a future direction of research.

Acknowledgments

We thank the reviewers for their careful reading of an earlier version of the article and for providing several helpful comments. We would also like to thank Purushottam Kar and Soham Sarkar for their valuable inputs which improved the article.

Bibliography

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., and Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, 60(1):4–19.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171.

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

Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*. Kendrick Press.

Chan, Y.-B. and Hall, P. (2009a). Robust nearest-neighbor methods for classifying high-dimensional data. *The Annals of Statistics*, 37(6A):3186–3203.

Chan, Y.-B. and Hall, P. (2009b). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, 96(2):469–478.

Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.

Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., and Batista, G. (2018). The UCR time series classification archive. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

Deegalla, S. and Bostrom, H. (2006). Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, pages 245–250. IEEE.

Dutta, S. and Ghosh, A. K. (2016). On some transformations of high dimension, low sample size data for nearest neighbor classification. *Machine Learning*, 102(1):57–83.

Francois, D., Wertz, V., and Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics: New York.

Globerson, A. and Roweis, S. (2005). Metric learning by collapsing classes. *Advances in Neural Information Processing Systems*, 18:451–458.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer, New York.

Li, Z. and Zhang, Y. (2020). On a projective ensemble approach to two sample test for equality of distributions. In *International Conference on Machine Learning*, pages 6020–6027. PMLR.

Thrampoulidis, C. (2020). Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Neural Information Processing Systems (NeuRIPS 2020)*.

Tomašev, N., Radovanović, M., Mladenić, D., and Ivanović, M. (2014). Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. *International Journal of Machine Learning and Cybernetics*, 5(3):445–458.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons.

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.

Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2).