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

Sarbojit Roy *Indian Institute of Technology Kanpur, Uttar Pradesh- 208016, India*

Jyotishka Ray Choudhury *Indian Statistical Institute, Kolkata, West Bengal - 700108, India*

Subhajit Dutta *Indian Institute of Technology Kanpur, Uttar Pradesh - 208016, India*

Abstract

In high dimension, low sample size (HDLSS) settings, the *distance concentration* phenomena affects the performance of several popular classifiers which are based on Euclidean distances. The high-dimensional behaviour of these classifiers is completely governed by the first and the 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 as well as 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 $\mathcal{X}_1 = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}\}$ and $\mathcal{X}_2 = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}\}$ are two sets of observations from $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively, and $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ is the training sample of size $n = n_1 + n_2$. Our objective is to classify a test point $\mathbf{Z} \in \mathbb{R}^p$ coming either from $\mathbf{F}_1$ or $\mathbf{F}_2$. The prior probability of $\mathbf{Z}$ coming from j-th class is given by $\mathrm{P}[\mathbf{Z} \sim \mathbf{F}_j] = \pi_j > 0$ with $\pi_1 + \pi_2 = 1$. We are interested in investigating the behavior of classification methods in high dimension, low sample size (HDLSS) settings, where the sample size n remains fixed, but the data dimension p increases.

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

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 be the mean vector and the covariance matrix of $\mathbf{F}_j$ for $j = 1, 2$. Let us 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 asymptotic 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 (Friedman et al., 2001), average distance (AVG) classi-

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

fier (Chan and Hall, 2009b), support vector machine (SVM) (Vapnik, 1998), etc. The authors showed that these methods are incapable of correctly classifyng an observation if the location difference between the competing populations is 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 $\sigma_1^2 \neq \sigma_2^2$. The existing 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 the ED, and the behavior of ED in the HDLSS asymptotic regime is completely governed by the constants ~~ν^2, σ_1^2 and σ_2^2~~. 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 applicable to ~~only~~ 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 classification methods, but under the assumption of gaussianity.

1.1 Motivation

Li and Zhang (2020) proposed a projective-ensemble based procedure for testing equality of two distributions. They considered a new measure of distance between $\mathbf{F}_1$ and $\mathbf{F}_2$ as defined below:

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

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

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

for $p \geq 1$. The authors showed that for a fixed p, $\tau = 0$ iff $\mathbf{F}_1 = \mathbf{F}_2$. This property of τ is particularly useful for distinguishing one distribution from another, and it can be utlized in classification as well. Note that τ is based on EDs and the inner product of p-dimensional vectors. Unsurprisingly, it exhibits an asymptotic behavior that is similar to the existing ED based methods in HDLSS setting.

To motivate the problem, we use τ and modify the scale adjusted average distance (SAVG) classifier (Chan and Hall, 2009b) by simply replacing the Euclidean norm $\|\mathbf{u}-\mathbf{v}\|^2$ with $h(\mathbf{u}, \mathbf{v})$ defined in Section 1.1. Our intention is to establish the fact that the resulting classifier (denoted by δ_0) is no different than its predecessors, i.e., it suffers from the same drawback that the other ED based methods have. To show the severity of *distance concentration* in high dimensions, let us 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 $k = 1, \ldots, p$. 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 performances of several classifiers, namely, 1NN, the usual SAVG, δ_0, SVM with linear kernel (SVM-LIN) and SVM with radial basis function (SVM-RBF) kernel. Details of the simulation study are given in Section 4.

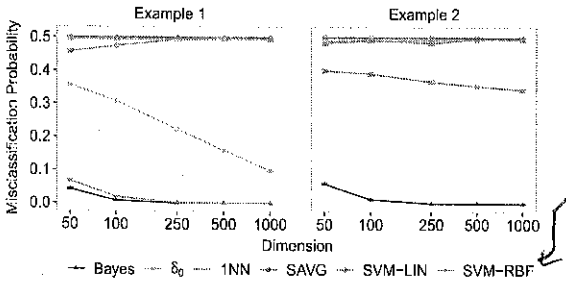

Figure 1: Performance of δ_0 and Popular Classifiers.

~~Observe that~~ In the first example, $\nu^2 = 0$ (since $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{1}_p$) but $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$, i.e., $\nu^2 < |\sigma_1^2 - \sigma_2^2|$. δ_0 uses the difference in scales and yields moderate performance. Whereas, the existing classifiers (except SVM-RBF) misclassify 50% of the observations. SVM-RBF, being a quadratic classifier performs well in this example since $\mathbf{F}_1$ and $\mathbf{F}_2$ have difference in their scales. The classification is a little 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 of the competing populations. As a result, δ_0 along with the existing classifiers fail to correctly classify the test observations. We will revisit these examples later in Section 4.

1.2 Our Contribution

In this article, we develop classifiers that are not limited by the existence of the constants ν^2, σ_1^2 and σ_2^2, or the relationship among them. 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 the two classifiers. Asymptotic properties of the classifiers are studied in Section 3. A theoretical result is presented to compare the performances of the classifiers in finite dimensions. Along with the HDLSS asymptotics, we also investigate the behavior of the proposed classifiers when both n and p increases. Performance of the classifiers is studied through a numerical exercise involving several simulated data sets in Section 4. We also investigate the behavior of the classifiers on some real data sets in Section 5. The article ends with concluding remarks in Section 6. All proofs, relevant mathematical details and codes are provided in the Supplementary.

2 METHODOLOGY

First, let us recall the classifier δ_0 mentioned in Section 1.1. For given training samples χ_1 and χ_2 with sizes $n_1(\geq 2)$ and $n_2(\geq 2)$, respectively and $\mathbf{z} \in \mathbb{R}^p$, the classifier δ_0 is formally defined as

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

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

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

Let Δ_0 be the misclassification probability of δ_0. For a random vector $\mathbf{Z}$ (independent from the training sample χ), Δ_0 is defined as $\mathrm{P}[\delta_0(\mathbf{Z}) \neq$ true class label of $\mathbf{Z}]$. 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 assume the following:

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

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

We also assume that the components of the sequence $\{U_kV_k, k \geq 1\}$ are weakly dependent, where $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$ for $j, j' \in \{1,2\}$. In particular,

(c) $\sum_{1\leq k<k'\leq p} \mathrm{Corr}(U_kV_k, U_{k'}V_{k'}) = o(p^2)$.

Assumption (c) is trivially satisfied if the component variables of the underlying populations are independently distributed. It continues to hold with some additional conditions on their dependence structure. For example, (c) is satisfied when the sequence $\{U_kV_k, k \geq 1\}$ has ρ-mixing property (Bradley, 2005; Hall et al., 2005). Conditions similar to (c) are frequently considered in the literature for studying high-dimensional behavior of various statistical procedures (Aoshima et al., 2018). We will revisit a similar condition later in Section 3.

Lemma 2.1 *If assumptions* (a)-(c) *are satisfied, then* $\sin(2\pi h(\mathbf{U},\mathbf{V}))$ *converges in probability to* $\nu_{jj'}/\left[(\sigma_j^2 + \nu_{jj})(\sigma_{j'}^2 + \nu_{j'j'})\right]^{\frac{1}{2}}$ *as* $p \to \infty$, *where* $\mathbf{U} \sim \mathbf{F}_j$ *and* $\mathbf{V} \sim \mathbf{F}_{j'}$ *for* $j, j' \in \{1,2\}$.

It is clear from Lemma 2.1 that the asymptotic behavior of a classifier based on h will be governed by the constants $\nu_{jj'}$, σ_j^2 for $j, j' \in \{1,2\}$. Theorem 2.2 below states that the classifier δ_0 is particularly useful in high dimensions when the underlying distributions differ either in their locations and/or scales.

Theorem 2.2 *Suppose that assumptions* (a)-(c) *are satisfied, and either of the following cases is true:*
(a) ν_{11}, ν_{12} *and* ν_{22} *are unequal (i.e.,* $\nu^2 > 0$*),*
(b) $\nu_{11} = \nu_{12} = \nu_{22} \neq 0$ *and* $\sigma_1^2 \neq \sigma_2^2$.
Then, for any $\pi_1 > 0$, $\Delta_0 \to 0$ *as* $p \to \infty$.

Recall Example 1, and note that it satisfies condition (b) in Theorem 2.2 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 δ_0 performed well in the first example, but failed in the second one. Now, we move ahead to develop classifiers whose asymptotic properties are not governed by the limiting constants. 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 between Two Distributions

If $\mathbf{X} = (X_1,\ldots,X_p)^\top \sim \mathbf{F}_1$ and $\mathbf{Y} = (Y_1,\ldots,Y_p)^\top \sim \mathbf{F}_2$ then we denote the distribution of X_k and Y_k by F_{1k} and F_{2k}, respectively,

for $k = 1, \ldots, p$. Suppose, $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} \mathbf{F}_1$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} \mathbf{F}_2$. Recall the definition of τ in Section 1.1 and note that the distance between F_{1k} and F_{2k} (obtained using the projective-ensemble approach) is given by $\tau_k = \mathrm{E}\big[h(X_{1k}, X_{2k}) - 2h(X_{1k}, Y_{1k}) + h(Y_{1k}, Y_{2k})\big]$. Here, $\tau_k \geq 0$ and equality holds iff $F_{1k} = F_{2k}$ for $k = 1, \ldots, p$. We denote the average of these distances as $\bar{\tau}_p = \sum_{k=1}^{p} \tau_k / p$. Clearly, $\bar{\tau}_p$ becomes 0 iff $\tau_k = 0$ for all $k = 1, \ldots, p$, i.e.,

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

This property of $\bar{\tau}_p$ suggests that it can be used as a *measure of separation* between $\mathbf{F}_1$ and $\mathbf{F}_2$. As long as the one-dimensional marginals of the competing populations are different, $\bar{\tau}_p$ is positive. This is the fundamental idea that we will use in developing new criteria for classification.

Recall the definition of h given in Section 1.1, and consider

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

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

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

For $\mathbf{z} \in \mathbb{R}^p$, we define the following:

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

$$\bar{T}_j(\mathbf{z}) = \frac{1}{n_j} \sum_{\mathbf{U} \in \mathcal{X}_j} \bar{h}_p(\mathbf{U}, \mathbf{z}) \text{ and}$$

$$\bar{L}_j(\mathbf{z}) = \bar{T}_{jj} - 2\bar{T}_j(\mathbf{z}) \text{ for } j = 1, 2. \tag{2.4}$$

Therefore,

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

Consequently, we get

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

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

Equation (2.6) shows the usefulness of $\bar{L}(\mathbf{Z}) = \bar{L}_2(\mathbf{Z}) - \bar{L}_1(\mathbf{Z})$ as a discriminant. For any $p \geq 1$, it is expected of $\bar{L}_2(\mathbf{Z})$ to be larger (resp., smaller) than $\bar{L}_1(\mathbf{Z})$ if $\mathbf{Z} \sim \mathbf{F}_1$ (resp., $\mathbf{Z} \sim \mathbf{F}_2$). Based on this observation, we propose the classifier:

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

for $\mathbf{z} \in \mathbb{R}^p$. One may also express $\delta_1(\mathbf{z})$ as $\delta_1(\mathbf{z}) = \arg\min_{j \in \{1,2\}} \bar{L}_j(\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^p$. The misclassification probability of δ_1 is given by $\Delta_1 = \mathrm{P}[\delta_1(\mathbf{Z}) \neq$ true label of $\mathbf{Z}]$. Note that δ_1 can be readily extended to deal with a multi-class classification problem. Let the number of classes be $J (\geq 2)$. For given random samples $\mathcal{X}_1, \ldots, \mathcal{X}_J$, we classify a new observation $\mathbf{Z}$ as $\delta_1(\mathbf{Z}) = \arg\min_{1 \leq j \leq J} \bar{L}_j(\mathbf{Z})$, where $\bar{L}_j(\mathbf{Z}), \bar{T}_j(\mathbf{Z})$ and $\bar{T}_{jj}$ are as defined in Eq. (2.4) for $j = 1, \ldots, J$.

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

The classifier δ_1 leverages $\bar{\tau}_p$, the average of distances between F_{1k} and F_{2k} for $k = 1, \ldots, p$, to classify a test point. However, the index $\bar{\tau}_p$ has some limitations. Consider

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

Since $\bar{\tau}_p \geq 0$, we always have $\bar{\tau}_p(1,2) \leq \{\bar{\tau}_p(1,1) + \bar{\tau}_p(2,2)\}/2$. Without loss of generality, let us assume that $\bar{\tau}_p(1,1) < \bar{\tau}_p(2,2)$. Now, if $\bar{\tau}_p(1,1) < \bar{\tau}_p(1,2) < \bar{\tau}_p(2,2)$, then $\{\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)\}$ and $\{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)\}$ are of different signs. Adding them up may nearly cancel each other, thus making the value of $\bar{\tau}_p$ close to 0, i.e., the distance between $\mathbf{F}_1$ and $\mathbf{F}_2$ may become seemingly negligible. One way to deal with this problem is to consider $\bar{\psi}_p = \{\bar{\tau}_p(1,1) - \bar{\tau}_p(1,2)\}^2 + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,2)\}^2$. This eliminates the possibility of such cancellations. It is easy to check that $\bar{\psi}_p$ has the following property:

$$\bar{\psi}_p = 0 \text{ iff } F_{1k} = F_{2k} \text{ for all } k = 1, \ldots, p.$$

Also, note that $\bar{\psi}_p$ can also be expressed as follows:

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

When $\bar{\tau}_p(1,2)$ lies between $\bar{\tau}_p(1,1)$ and $\bar{\tau}_p(2,2)$, $\bar{\psi}_p$ amplifies the measure of separability between $\mathbf{F}_1$ and $\mathbf{F}_2$. Under such circumstances, a classifier that utilizes this amplified measure $\bar{\psi}_p$ is shown to have better classification accuracy than δ_1. The modification

given in (2.8) is similar to what Biswas and Ghosh (2014) suggested for improving the power of energy based tests.

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

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

We intend to construct a discriminant that converges to $\bar{\psi}_p$ (respectively, $-\bar{\psi}_p$) as $p \to \infty$ if $\mathbf{Z} \sim \mathbf{F}_1$ (respectively, $\mathbf{Z} \sim \mathbf{F}_2$). For a given $\mathbf{z} \in \mathbb{R}^p$, we define

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

We show that $|\bar{\theta}(\mathbf{Z})|$ is a consistent estimator of $\bar{\psi}_p$. In particular, we will prove that $\bar{\theta}(\mathbf{Z})$ converges in probability to $\bar{\psi}_p$ if $\mathbf{Z} \sim \mathbf{F}_1$ and to $-\bar{\psi}_p$ if $\mathbf{Z} \sim \mathbf{F}_2$ (see Section 3 for details). This motivates us to define the following classifier for $\mathbf{z} \in \mathbb{R}^p$:

$$\delta_2(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{\theta}(\mathbf{z}) > 0, \\ 2, & \text{otherwise.} \end{cases} \tag{2.10}$$

Let Δ_2 be the misclassfication probability of δ_2, i.e., $\Delta_2 = \mathrm{P}[\delta_2(\mathbf{Z}) \neq \text{ true label of } \mathbf{Z}]$. Unlike the classifier δ_1, δ_2 cannot be readily extended to deal with J-class problems with $J > 2$. We implement the idea of 'majority voting' (Friedman et al., 2001) to obtain the misclassification rate of δ_2 when $J > 2$.

> Computing $\bar{T}_{jj'}$ and $\bar{T}_j(\mathbf{z})$ for $\mathbf{z} \in \mathbb{R}^p$ require $O(n^2 p)$ and $O(np)$ operations, respectively, for $j, j' \in \{1, \ldots, J\}$. Overall complexity is $O(n^2 p)$. It increases linearly with respect to p and makes the methods advantageous in analyzing high-dimensional data sets.

Recall Examples 1 and 2 introduced in Section 1. Figure 2 shows that the classifier δ_2 has substantial improvement over δ_1 in terms of misclassfication probability in both the examples. This improvement is due to the fact that $\bar{T}_{12}$ lies between $\bar{T}_{11}$ and $\bar{T}_{22}$

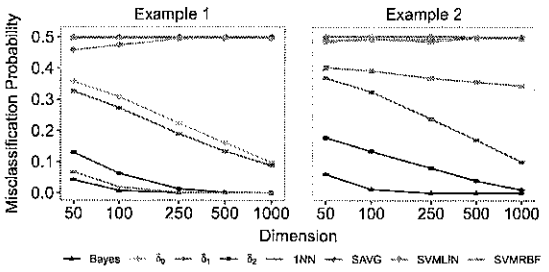

Figure 2: Performance of the Proposed Classifiers.

in both examples (see Table S1 in the Supplementary). A theoretical explanation of such behavior of δ_1 and δ_2 is presented in Section 3.1.2. Examples 1 and 2 establish the necessity of the modified distance measure in Eq. (2.8) and the advantage of using δ_2 over δ_1 in such scenarios.

3 ASYMPTOTIC PROPERTIES

Let us now investigate the asymptotic behavior of the classifiers δ_1 and δ_2 in both HDLSS and *ultrahigh-dimensional* settings. Note that in the HDLSS settings, n is fixed and $p \to \infty$, whereas in the *ultrahigh-dimensional* asymptotic regime, p may grow simultaneously with n. First, we show that under fairly general conditions, the proposed classifiers yield *perfect classification* in HDLSS settings.

3.1 Asymptotic Behavior in HDLSS Settings

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

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

Although A1 is similar to (c) introduced in Section 2, it is a weaker assumption than (c) because (c) requires the first and second order moments to exist. Such moment conditions are not needed for A1 due to boundedness of the function h.

Similar to (c), A1 is trivially satisfied if the component variables of the underlying populations 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 the ρ-mixing property. Note that if the sequence of component variables

$\{U_k, k \geq 1\}$ and $\{V_k, k \geq 1\}$ have ρ-mixing property, then $\{h(U_k, V_k), k \geq 1\}$ has ρ-mixing property for every measurable function h (see Bradley, 2007, Theorem 6.6-II). The next result shows that assumption A1 is a sufficient condition for convergence of the discriminants $\bar{L}(\mathbf{Z})$ and $\bar{\theta}(\mathbf{Z})$, $\mathbf{Z} \in \mathbb{R}^p$.

Lemma 3.1 *If* A1 ~~*is satisfied*~~, *then for a test observation* $\mathbf{Z} \in \mathbb{R}^p$, *we have*

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

(b) If $\mathbf{Z} \sim \mathbf{F}_2$, *then* $|\bar{L}(\mathbf{Z}) + \bar{\tau}_p| \xrightarrow{\mathrm{P}} 0$ *and* $|\bar{\theta}(\mathbf{Z}) + \bar{\psi}_p| \xrightarrow{\mathrm{P}} 0$ *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 $\mathbf{Z} \sim \mathbf{F}_2$. We have seen that $\bar{\tau}_p = \bar{\psi}_p = 0$ iff $F_{1k} = F_{2k}$ for all $k = 1, \ldots, p$. 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 A2 is satisfied if the component variables of $\mathbf{U} \sim \mathbf{F}_j$ are identically distributed for $j = 1, 2$. Then, $\tau_k = \tau_1$ for all $k \geq 1$, making $\bar{\tau}_p (= \tau_1)$ is free of p and positive. It follows from the definition of $\bar{\psi}_p$ in Eq. (2.8) that A2 implies $\liminf_p \bar{\psi}_p > 0$.

3.1.1 Asymptotic Behavior of δ_1 and δ_2

Now, we discuss properties of the classifier δ_1 in HDLSS setting. If $\bar{\tau}_p$ does not vanish with increasing dimension, then Lemma 3.1 suggests that the random variable $\bar{L}(\mathbf{Z})$ converges to a positive (respectively, negative) value where $\mathbf{Z} \sim \mathbf{F}_1$ (respectively, $\mathbf{F}_2$). The following theorem states that under fairly general conditions, the proposed classifier δ_1 perfectly classifies an observation as the dimension increases.

Theorem 3.2 *If* A1 *and* A2 *are satisfied, then for any* $\pi_1 > 0$, *the misclassfication probability of* δ_1 *converges to 0 as* $p \to \infty$.

Observe that the asymptotic behavior of the classifier δ_1 is no longer governed by the constants $\nu_{jj'}$, σ_j^2 for $j, j' \in \{1, 2\}$. In fact, the behavior does not depend on the existence of moments. In this sense, the classifier δ_1 is clearly robust.

The second classifier δ_2 has asymptotic properties similar to the classifier δ_1. Under the same set of assumptions on the distributions, δ_2 yields *perfect classification* as $p \to \infty$.

Theorem 3.3 *If* A1 *and* A2 *are satisfied, then for any* $\pi_1 > 0$, *the misclassfication probability of* δ_2 *converges to 0 as* $p \to \infty$.

> The asymptotic behavior of the classifiers is free of moment conditions.
>
> They perfectly classify an observation if $\mathbf{F}_1$ and $\mathbf{F}_2$ have asymptotically non-zero separation and components are weakly dependent.

One should observe that assumptions A1 and A2 are fairly general, and Theorems 3.2 and 3.3 are stronger than what exist in the 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, i.e., A1 and A2. The next result provides a sufficient condition under which one classifier performs better than the other in finite dimensions. 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 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, and the other is negative. As a result, $\bar{\tau}_p$ may take a small value (see the discussion in Section 2.2). The next result suggests that under such circumstances, δ_2 has better performance than δ_1.

Theorem 3.4 *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 in Examples 1-2, $\bar{T}_{12}$ lies between $\bar{T}_{11}$ and $\bar{T}_{22}$ (see Table S1 in Supplementary). Thus, A3

is satisfied in both the examples. As a result, we see δ_2 yielding lower misclassfication probabilities than δ_1 (see Figure 2).

3.2 Asymptotic Behavior of δ_1 and δ_2 when Sample Size Increases

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

A4. There exists a $\beta \in [0, 1)$ 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 where $\log p_n = O(n^\beta)$ with $\beta = 0$. We also assume that $\lim_{n\to\infty} n_1/n = \pi_1 > 0$.

Let us first present the 'oracle' versions of the classifiers δ_1 and δ_2. If $\mathbf{F}_1$ and $\mathbf{F}_2$ are known, then the 'oracle' version of δ_1 classifies a vector $\mathbf{z}$ as following:

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

where $\bar{L}^0(\mathbf{z}) = \bar{L}_2^0(\mathbf{z}) - \bar{L}_1^0(\mathbf{z})$, with $\bar{L}_j^0(\mathbf{Z}) = \bar{\tau}_p(j,j) - 2E[h_p(\mathbf{U}, \mathbf{Z}) \mid \mathbf{Z}]$ for $\mathbf{U} \sim \mathbf{F}_j$, $j = 1, 2$. We call it an 'oracle' classifier since the underlying distributions are completely known, and the classification method does not depend on the training data $\mathcal{X}$. 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 $2\bar{\theta}^0(\mathbf{z}) = \bar{\tau}_p \bar{L}^0(\mathbf{z}) + \{\bar{\tau}_p(2,2) - \bar{\tau}_p(1,1)\} \times \{\bar{L}_2^0(\mathbf{z}) + \bar{L}_1^0(\mathbf{z}) + 2\bar{\tau}_p(1,2)\}$. The misclassification probability of δ_j^0 is given by $\Delta_j^0 = P[\delta_j^0(\mathbf{Z}) \neq \text{true label of } \mathbf{Z} \mid j = 1, 2$. Note that $\bar{L}(\mathbf{z})$ and $\bar{\theta}(\mathbf{z})$ are in fact estimators of $\bar{L}^0(\mathbf{z})$ and $\bar{\theta}^0(\mathbf{z})$, respectively. Therefore, δ_j is an estimator of δ_j^0 for $j = 1, 2$.

In this section, we show that under absolute continuity of the competing distributions, Δ_j converges to Δ_j^0 for $j = 1, 2$ as $n \to \infty$. Furthermore, we derive an upper bound on the rate of the convergence. First, we look in to the convergence of the discriminants $\bar{L}(\mathbf{z})$ and $\bar{\theta}(\mathbf{z})$ for $\mathbf{z} \in \mathbb{R}^{p_n}$.

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

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

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

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

Since $1 - 2\gamma > \beta$, $e^{-\{n^{1-2\gamma} - n^\beta\}} = o(1)$ as $n \to \infty$. The above result shows that $\bar{L}(\mathbf{z})$ and $\bar{\theta}(\mathbf{z})$ converge to $\bar{L}^0(\mathbf{z})$ and $\bar{\theta}^0(\mathbf{z})$, respectively, at an exponential rate as n increases. We use Hoeffding's inequality and the bounded difference inequality (Wainwright, 2019) to derive the rates. As a consequence of Lemma 3.5, we have the next result.

Theorem 3.6 *Suppose assumption* A4 *is satisfied for some $0 \le \beta < 1$. Fix $0 < \gamma < (1-\beta)/2$. Then, for any $\pi_1 > 0$, there exist positive constants B_0 and B_1 such that*

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

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

Corollary 3.7 *If assumption* A4 *is satisfied with $\beta = 0$ and $\mathbf{F}_1, \mathbf{F}_2$ are absolutely continuous, then $\Delta_1 - \Delta_1^0$ and $\Delta_2 - \Delta_2^0$ converge to 0 as $n \to \infty$.*

4 SIMULATION STUDY

In this section, we analyze some simulated data sets to compare the performance of δ_0, δ_1 and δ_2 with some well-known classifiers, namely, GLMNET (Friedman et al., 2001), 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 R statistical software. Details about the packages used and other parameters of the simulation exercise are provided in the Supplementary.

Recall Examples 1 and 2 introduced in Section 1. Three additional examples are considered to demonstrate the performance of the proposed 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 $Y_{1k} \overset{i.i.d.}{\sim} C(0,2)$,

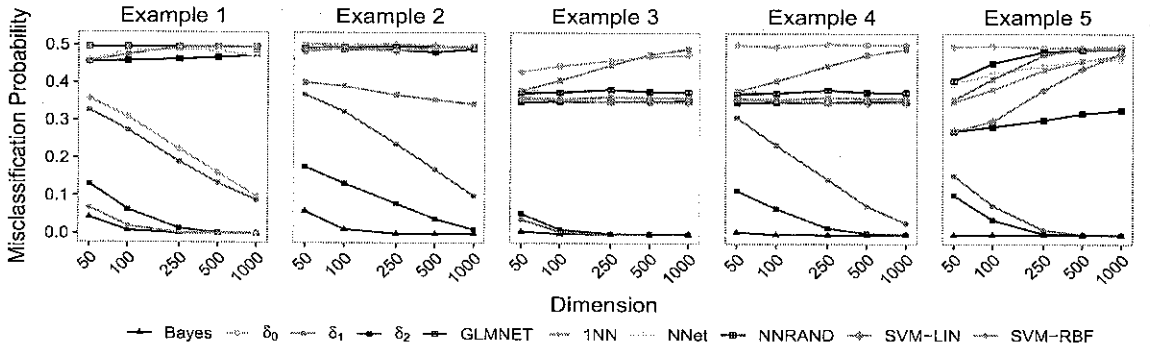

Figure 3: Average Misclassification Rates of Classifiers for Fixed n and Varying p.

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 $k = 1, \ldots, p$. 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 are a location, scale and location-scale problem, respectively. All three examples involve heavy-tailed distributions. Note that $\mathbf{F}_1$ and $\mathbf{F}_2$ in Examples 1-4 are symmetric while in Example 5 they are positively-skewed. 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 compute the average misclassification rates, which are reported in Figure 3 along with the standard errors.

First of all, observe that assumption A1 is satisfied for Examples 1-5 since the components variables in these examples are independently distributed. Also, the components have identical marginals. Therefore, $\bar{\tau}_p (= \tau_1 > 0)$ is free of both n and p. Hence, A2 is satisfied. Thus, Theorems 3.2 and 3.3 hold, and the proposed classifiers yield promising results for all five examples.

It is clear from Figure 3 that both Δ_1 and Δ_2 decreases to 0 in all examples. However, $\Delta_1 \leq \Delta_2$ in Example 3, while $\Delta_2 \leq \Delta_1$ in Examples 1, 2, 4 and 5. Note that $\bar{T}_{12} \leq \min\{\bar{T}_{11}, \bar{T}_{22}\}$ in Example 3 and $\bar{T}_{12} \geq \min\{\bar{T}_{11}, \bar{T}_{22}\}$ in other examples. The estimated values for each example are provided in Table S1 of Supplementary. Earlier in Theorem 3.4, we have discussed how the relationship among $\bar{\tau}_p(1,1), \bar{\tau}_p(1,2)$ and $\bar{\tau}_p(2,2)$ determine the ordering between Δ_1 and Δ_2.

The existing classifiers exhibit poor performance in general, except for a few occasions. We observe that SVM-RBF performs well in Example 1 since $\mathbf{F}_1$ and $\mathbf{F}_2$ have difference in scales. The rest of the methods (including δ_0), fail miserably due to their inability to look beyond difference in locations in high dimensions. In Examples 3, 4 and 5, the existing classifiers could not perform well since the moments of the competing distributions do not exist.

Recall that the computational complexity of the proposed classifiers is $O(n^2 p)$. We report average time taken by these classifiers to classify a test observation in Table S2 of the Supplementary. It clearly shows the advantage of using δ_1 and δ_2 over the popular classifiers.

5 REAL DATA ANALYSIS

In this section, 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 a *fixed training set and a fixed test set*. For our analysis, we combined the available training and test data, and 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 other half was considered as the test set. This procedure was repeated 100 times to obtain stable estimates of the misclassification probabilities.

The `Computers` data contains readings on electricity consumption from 251 households in UK, sampled in two-minute intervals over a month. Each data point is of length 720. Classes are Desktop and Laptop devices with 250 (125 training and 125 test) samples in each. From Table 1, we observe that $\Delta_2 < \Delta_1$ for this data. For this data set, the estimates of $\bar{\tau}_p(1,1)$,

$\tau_p(1,2)$, and $\tau_p(2,2)$ are given by $\bar{T}_{11} = 0.972$, $\bar{T}_{12} = 1.043$, and $\bar{T}_{22} = 1.155$, respectively. Observe that $\bar{T}_{12}$ lies between $\bar{T}_{11}$ and $\bar{T}_{22}$. Hence, δ_2 dominates δ_1. In fact, δ_2 produces the lowest misclassification rate of all. Among the existing classifiers, GLMNET and SVM-RBF are better than the others, securing the third and fourth positions, respectively. We also note that the ED between the sample means (when scaled by dimension) of the two classes attains the value 7.75×10^{-11}. It can be safely assumed that the two classes have no difference in their locations. This explains poor performance of some of the linear classifiers, e.g., SVM-LIN.

The second data set `SmoothSubspace` has observations for testing whether a clustering algorithm is able to extract smooth subspaces for clustering time series data. Each time series is of length 15. The 3 classes having 100 (50 train and 50 test) observations each, correspond to which cluster the time series belongs to. For this data set, we observe in Table 1 that $\Delta_1 < \Delta_2$. Note that in this data set, $\bar{T}_{jj'}$ is less than both $\bar{T}_{jj}$ and $\bar{T}_{j'j'}$ for all $1 \le j \ne j' \le 3$. The estimated values are given as $\bar{T}_{11} = 1.384$, $\bar{T}_{22} = 1.378$, $\bar{T}_{33} = 1.386$, $\bar{T}_{12} = 1.340$, $\bar{T}_{13} = 1.326$, and $\bar{T}_{23} = 1.314$. Among the existing classifiers, SVM-RBF yields the lowest misclassification rate, 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)
S.Sub.	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)

Sec Cui et al. (2015) for other approaches of robust classification in the presence of sparsity. Nevertheless, analysis of several simulated and real data sets show the promising performance of the classifiers.

6 CONCLUDING REMARKS

In this article, we present classifiers capable of perfectly classifying a test observation if the one-dimensional marginals of the underlying distributions are different. We study the theoretical properties of the classifiers in the HDLSS and *ultra high-dimensional* setting. The proposed classifiers are robust, i.e., they yield *perfect classification* even when the competing distributions are heavy-tailed. Their tuning free classification criteria make the methods fast and easy to implement. However, these classifiers may fail if the one-dimensional marginals are identical, or the component variables with differences in their marginals are sparsely located. Observe that under such circumstances, assumption A2 is not satisfied, i.e., $\liminf_p \bar{\tau}_p = 0$.

Acknowledgments

We would like to thank the associate editor and the five anonymous reviewers for their careful reading of an earlier version of the article and for providing us with helpful comments. We would also like to thank Soham Sarkar for his valuable inputs which helped to improve 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/`.

Dedieu, A. (2019). Error bounds for sparse classifiers in high-dimensions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 48–56. PMLR.

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.

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).