



On High-Dimensional Modifications of the Nearest Neighbor Classifier

Annesha Ghosh¹, Deep Ghoshal², Bilol Banerjee¹ and Anil K. Ghosh¹

¹*Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India*

²*Department of Statistics, University of Illinois at Urbana-Champaign, USA.*

Received: 06 May 2024; Revised: 26 September 2024; Accepted: 30 September 2024

Abstract

Nearest neighbor classifier is arguably the most simple and popular nonparametric classifier available in the literature. However, due to the concentration of pairwise distances and the violation of the neighborhood structure, this classifier often suffers in high-dimension, low-sample size (HDLSS) situations, especially when the scale difference between the competing classes dominates their location difference. Several attempts have been made in the literature to take care of this problem. In this article, we discuss some of those existing methods and propose some new ones. We carry out some theoretical investigations in this regard and analyze several simulated and benchmark datasets to compare the empirical performances of our proposed methods with some of the existing ones.

Key words: Dimension reduction; Feature extraction; HDLSS asymptotics; Mixture distributions; Nearest neighbors.

AMS Subject Classifications: 62H30, 68T10

1. Introduction

In supervised classification, we use a training set of labeled observations from different competing classes to form a decision rule for classifying unlabeled test set observations as accurately as possible. Starting from Fisher (1936), Rao (1948) and Fix and Hodges (1951), several parametric as well as nonparametric classifiers have been developed for this purpose (see, *e.g.*, Duda *et al.*, 2007; Hastie *et al.*, 2009). Among them, the nearest neighbor classifier (see, *e.g.*, Cover and Hart, 1967) is perhaps the most popular one. The k -nearest neighbor classifier (k -NN) classifies an observation \mathbf{x} to the class having the maximum number of representatives among the k nearest neighbors of \mathbf{x} . This classifier works well if the training sample size is large compared to the dimension of the data. For a suitable choice of k (which increases with the training sample size at an appropriate rate), under some mild regularity conditions, the misclassification rate of the k -NN classifier converges to the Bayes risk (*i.e.*, the misclassification rate of the Bayes classifier) as the training sample size grows to infinity (see, *e.g.* Devroye *et al.*, 2013; Hall *et al.*, 2008). However, like other nonparametric

methods, this classifier also suffers from the curse of dimensionality (see, *e.g.*, Carrerira-Perpinan, 2009), especially when the dimension of the data is much larger than the training sample size. In such high-dimension, low-sample-size (HDLSS) situations, the concentration of pairwise distances (see, *e.g.*, Hall *et al.*, 2005; François *et al.*, 2007), presence of hubs and the violation of the neighborhood structure (see, *e.g.*, Radovanovic *et al.*, 2010; Pal *et al.*, 2016) often have adverse effects on the performance of the nearest neighbor classifier.

To demonstrate this, we consider some simple examples involving two d -dimensional normal distributions. Descriptions of these examples are given below.

Examples 1 - 3: *In these three examples, the first class has a normal distribution with the mean vector $\mathbf{0}_d = (0, 0, \dots, 0)^\top$ and the dispersion matrix \mathbf{I}_d (the $d \times d$ identity matrix), while the second class has the mean vector $\mu\mathbf{1}_d = \mu(1, 1, \dots, 1)^\top = (\mu, \mu, \dots, \mu)^\top$ and the dispersion matrix $\sigma^2\mathbf{I}_d$. In Example 1, we consider a location problem where we take $\mu = 1$ and $\sigma = 1$. Example 2 deals with a location-scale problem with $\mu = 1$ and $\sigma = 2$. As Example 3, we choose a scale problem, where μ and σ are taken as 0 and 2, respectively.*

In each of these examples, we carry out our experiment for 7 different choices of d ranging between 10 and 1000 ($d = 10, 20, 50, 100, 200, 500$ and 1000). In each case, taking an equal number of observations from the two competing classes, we form the training and test sets of size 50 and 500, respectively. This is done 100 times, and the average test set misclassification rates of the 1-NN classifier over these 100 trials are reported in Figure 1.

Note that in each of these examples, the distribution of each measurement variable differs in two competing classes. So, each of them contains information about class separability, and as a result, the separability between the two classes increases with the dimension. One can check that in each of these examples, the Bayes risk converges to 0 as the dimension grows. Therefore, the misclassification rate of any good classifier is also expected to go down as the dimension increases. We observed the same for the 1-NN classifier in Example 1 (location problem), but surprisingly, in the other two cases, its misclassification rates were close to 0.5 in high dimensions.

A careful investigation explains the reasons for this diametrically opposite behavior. Let $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{n_1}\}$ and $\{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{n_2}\}$ be the training samples from two competing classes (here we have $n_1 = n_2 = 25$) $N(\mathbf{0}_d, \mathbf{I}_d)$ and $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, respectively. Now, for a test case \mathbf{Z} from $N(\mathbf{0}_d, \mathbf{I}_d)$, one can show that for each $i = 1, 2, \dots, n_1$, $\frac{1}{d}\|\mathbf{Z} - \mathbf{X}_i\|^2$, being the average of independent and identically distributed (i.i.d.) random variables, converges in probability to 2 as d increases to infinity. Similarly, it can be shown that for each $i = 1, 2, \dots, n_2$, $\frac{1}{d}\|\mathbf{Z} - \mathbf{Y}_i\|^2 \xrightarrow{P} 1 + \mu^2 + \sigma^2$. So, \mathbf{Z} is correctly classified by the 1-NN classifier (or any k -NN classifier with $k \leq \min\{n_1, n_2\}$) if $\mu^2 + \sigma^2 > 1$. Note that it was the case in all three examples. So, all observations from $N(\mathbf{0}_d, \mathbf{I}_d)$ were correctly classified. But for a test case \mathbf{Z}' from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, we have $\frac{1}{d}\|\mathbf{Z}' - \mathbf{X}_i\|^2 \xrightarrow{P} 1 + \mu^2 + \sigma^2$ for $i = 1, 2, \dots, n_1$ and $\frac{1}{d}\|\mathbf{Z}' - \mathbf{Y}_i\|^2 \xrightarrow{P} 2\sigma^2$ for $i = 1, 2, \dots, n_2$. So, it is correctly classified if and only if $\sigma^2 < 1 + \mu^2$. This condition was satisfied in Example 1, but not in the other two cases. Because of this violation of the neighborhood structure (where observations from one class have all neighbors from other classes), in Examples 2 and 3, the 1-NN classifier misclassified all observations from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$ and had misclassification rates close to 0.5.

This phenomenon of distance concentration in high dimension was observed by Hall *et al.* (2005) for Euclidean distances and François *et al.* (2007) for fractional distances. Hall *et al.* (2005) also studied the high dimensional behavior of some popular classifiers and observed this undesirable behavior of the nearest neighbor classifier. To take care of this problem, Chan and Hall (2009b) proposed an adjustment for the scale difference between the competing classes. They suggested to compute

$$\rho_1(\mathbf{Z}, \mathbf{X}_i) = \|\mathbf{Z} - \mathbf{X}_i\|^2 - \frac{1}{2} \binom{n_1}{2}^{-1} \sum_{s < t} \|\mathbf{X}_s - \mathbf{X}_t\|^2 \text{ for } i = 1, 2, \dots, n_1,$$

$$\rho_2(\mathbf{Z}, \mathbf{Y}_i) = \|\mathbf{Z} - \mathbf{Y}_i\|^2 - \frac{1}{2} \binom{n_2}{2}^{-1} \sum_{s < t} \|\mathbf{Y}_s - \mathbf{Y}_t\|^2 \text{ for } i = 1, 2, \dots, n_2$$

and classify \mathbf{Z} to the first (respectively, second) class if $\min \rho_1(\mathbf{Z}, \mathbf{X}_i) < \min \rho_2(\mathbf{Z}, \mathbf{Y}_i)$ (respectively, $\min \rho_1(\mathbf{Z}, \mathbf{X}_i) > \min \rho_2(\mathbf{Z}, \mathbf{Y}_i)$). Note that without the scale adjustments (second terms on the right-hand side of the equations), it turns out to be the usual 1-NN classifier. Figure 1 also shows the performance of this classifier (we refer to it as the CH classifier) in Examples 1-3. In Example 1, it performed like the 1-NN classifier. Interestingly, in Example 2, while the 1-NN classifier failed, this scale adjustment led to improved performance by the CH classifier in high dimensions. But in Example 3, like the 1-NN classifier, it also misclassified almost 50% observations. Note that for any \mathbf{Z} from $N(\mathbf{0}_d, \mathbf{I}_d)$, here $\rho_1(\mathbf{Z}, \mathbf{X}_i)/d \xrightarrow{P} 1$ ($i = 1, 2, \dots, n_1$) and $\rho_2(\mathbf{Z}, \mathbf{Y}_i)/d \xrightarrow{P} 1 + \mu^2$ ($i = 1, 2, \dots, n_2$) as d increases. So, it is correctly classified if $\mu^2 > 0$. Again, for any \mathbf{Z}' from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, we have $\rho_1(\mathbf{Z}', \mathbf{X}_i)/d \xrightarrow{P} \mu^2 + \sigma^2$ for $i = 1, 2, \dots, n_1$ and $\rho_2(\mathbf{Z}', \mathbf{Y}_i)/d \xrightarrow{P} \sigma^2$ for $i = 1, 2, \dots, n_2$. So, here also, we need $\mu^2 > 0$ for correct classification. In Examples 1 and 2, we had $\mu^2 > 0$. So, the CH classifier performed well in those two examples for large values of d . But in Example 3, where we had $\mu^2 = 0$, it misclassified almost 50% observations. This example shows that the CH classifier may fail to discriminate between two high-dimensional distributions differing only in their scales.

However, if we slightly modify Chan and Hall (2009b)'s proposal of scale adjustment, we can take care of high dimensional scale problems as well. Our modified version (which

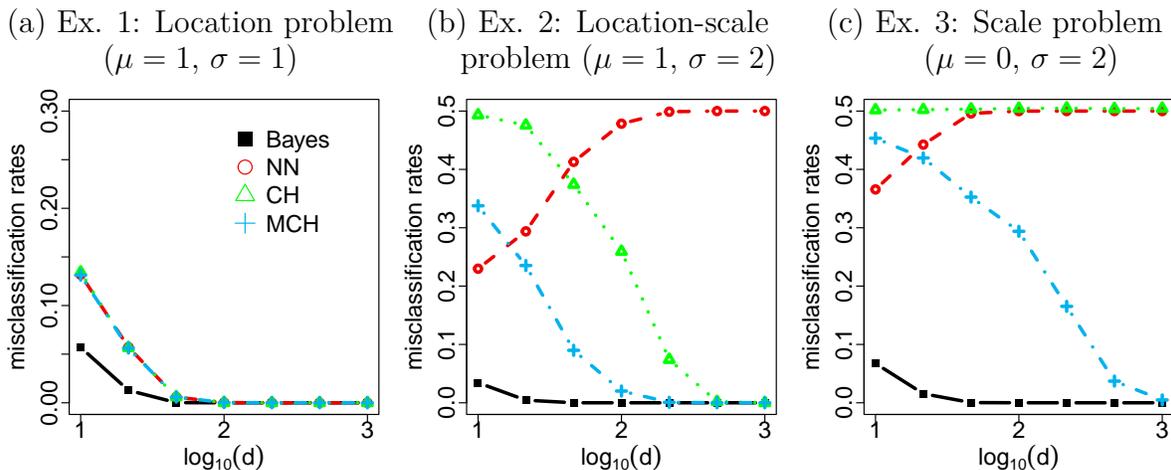


Figure 1: Misclassification rates of Bayes, NN, CH and MCH classifiers in Examples 1-3.

we refer to as the Modified Chan and Hall classifier or the MCH classifier) had excellent performance in all three examples (see Figure 1), especially for large values of d . In Section 2, we propose this modification and carry out some theoretical and numerical studies to understand the high-dimensional behavior of the resulting classifier.

An alternative strategy to deal with any high-dimensional problem is to reduce the dimension of the data and work on the reduced subspace. The simplest method of dimension reduction is to consider some random linear projections (see, *e.g.*, Fern and Brodley, 2003; Fradkin and Madigan, 2003; Vempala, 2005, and the references therein), and we can adopt that method for nearest neighbor classification as well. Another popular approach is to use projections based on principal component analysis (see, *e.g.*, Deegalla and Bostrom, 2006; Maciończyk *et al.*, 2023). But as pointed out in Dutta and Ghosh (2016), these methods often lead to poor performance in high-dimensional classification problems. For instance, from our description, it is quite clear that neither the principle component directions nor the random projections are meaningful in Example 3. Also getting consistent estimates of the principal components in high dimension is challenging (see, *e.g.*, Jung and Marron, 2009). Other approaches towards nearest neighbor classification of high-dimensional data include those based on mean absolute difference of distances (see, *e.g.*, Pal *et al.*, 2016; Roy *et al.*, 2022), hubness-based fuzzy measures (see, *e.g.*, Tomašev *et al.*, 2014) and distance metrics learning (see, *e.g.* Weinberger and Saul, 2009). Chan and Hall (2009a) proposed a robust version of the nearest neighbor method for classifying high-dimensional data, but their method can be used only for a specific type of two-class location problem. Instead of using random projections or principal components, Dutta and Ghosh (2016) suggested extracting some distance-based features from the data and performing nearest neighbor classification based on those features. They proposed two such methods, one using transformation based on average distances (TRAD) and the other using transformation based on inter-point distances (TRIPD). We briefly discuss these two methods in Section 3 and also propose some other methods for selecting distance-based features for nearest neighbor classification of high dimensional data. A comparative discussion of these methods is also given in this section based on our analysis of some simulated data sets. Some benchmark data sets are analyzed in Section 4 to compare the performances of these methods with two popular state-of-the-art classifiers, support vector machines (see, *e.g.* Cristianini and Shawe-Taylor, 2003; Steinwart and Christmann, 2008; Scholkopf and Smola, 2018) and random forest (see, *e.g.*, Breiman, 2001; Genuer and Poggi, 2020), which are known to perform well for high dimensional data. Finally, a brief summary of the work and some concluding remarks are given in Section 5. All proofs and mathematical details are given in the Appendix.

2. Modified scale-adjusted nearest neighbor classifier

We have seen that while the 1-NN classifier failed in Examples 2 and 3, the scale-adjusted CH classifier worked well in Example 2 when the dimension was large. But, in Example 3, this scale adjustment could not improve the performance of the 1-NN classifier.

This motivates us to look for a modified scale adjustment. We define

$$\rho_1^*(\mathbf{Z}, \mathbf{X}_i) = \|\mathbf{Z} - \mathbf{X}_i\| - \frac{1}{2} \binom{n_1}{2}^{-1} \sum_{s < t} \|\mathbf{X}_s - \mathbf{X}_t\| \text{ for } i = 1, 2, \dots, n_1,$$

$$\rho_2^*(\mathbf{Z}, \mathbf{Y}_i) = \|\mathbf{Z} - \mathbf{Y}_i\| - \frac{1}{2} \binom{n_2}{2}^{-1} \sum_{s < t} \|\mathbf{Y}_s - \mathbf{Y}_t\| \text{ for } i = 1, 2, \dots, n_2$$

and classify a test set observation \mathbf{Z} to the first (respectively, second) class if $\min \rho_1^*(\mathbf{Z}, \mathbf{X}_i)$ is smaller (respectively, larger) than $\min \rho_2^*(\mathbf{Z}, \mathbf{Y}_i)$. Figure 1 shows that this modified scale adjusted nearest neighbor classifier (henceforth referred to as the Modified Chan and Hall classifier or the MCH classifier) had excellent performance in high dimensions in all three examples. A small theoretical analysis explains the reasons for its superior performance.

Following our previous discussion on distance convergence, one can show that for a test set observation \mathbf{Z} from $N(\mathbf{0}_d, \mathbf{I}_d)$, as d tends to infinity, we have $\rho_1^*(\mathbf{Z}, \mathbf{X}_i)/\sqrt{d} \xrightarrow{P} 1/\sqrt{2}$ for $i = 1, 2, \dots, n_1$, while $\rho_2^*(\mathbf{Z}, \mathbf{Y}_i)/\sqrt{d} \xrightarrow{P} \sqrt{1 + \mu^2 + \sigma^2} - \sigma/\sqrt{2}$ for $i = 1, 2, \dots, n_2$. So, it is correctly classified if $\sqrt{1 + \mu^2 + \sigma^2} > (\sigma + 1)/\sqrt{2} \Leftrightarrow 1 + \mu^2 + \sigma^2 > (\sigma + 1)^2/2 \Leftrightarrow \mu^2 + \frac{1}{2}(\sigma - 1)^2 > 0$. Again for an observation \mathbf{Z}' from $N(\mu\mathbf{1}_d, \sigma^2\mathbf{I}_d)$, as $d \rightarrow \infty$, we have $\rho_1^*(\mathbf{Z}', \mathbf{X}_i)/\sqrt{d} \xrightarrow{P} \sqrt{1 + \mu^2 + \sigma^2} - 1/\sqrt{2}$ for $i = 1, 2, \dots, n_1$ and $\rho_2^*(\mathbf{Z}', \mathbf{Y}_i)/\sqrt{d} \xrightarrow{P} \sigma/\sqrt{2}$ for $i = 1, 2, \dots, n_2$. So, here also, \mathbf{Z}' is correctly classified if $\mu^2 + \frac{1}{2}(\sigma - 1)^2 > 0$. This inequality holds in all three examples considered in Section 1. This was the reason for the excellent performance of the MCH classifier in high dimensions.

Like the usual nearest neighbor classifier, multi-class generalizations of CH and MCH classifiers are quite straightforward. If there are J competing classes F_1, F_2, \dots, F_J with the training samples $\{\mathbf{X}_{j1}, \mathbf{X}_{j2}, \dots, \mathbf{X}_{jn_j}\}$ from the j -th class, ($j = 1, 2, \dots, J$), for classifying a test case \mathbf{Z} by the CH classifier, we can compute

$$\rho_j(\mathbf{Z}, \mathbf{X}_{ji}) = \|\mathbf{Z} - \mathbf{X}_{ji}\|^2 - \frac{1}{2} \binom{n_j}{2}^{-1} \sum_{s < t} \|\mathbf{X}_{js} - \mathbf{X}_{jt}\|^2 \text{ for } j = 1, 2, \dots, J, \ i = 1, 2, \dots, n_j$$

and assign \mathbf{Z} to class j_0 if $\min_{1 \leq i \leq n_{j_0}} \rho_{j_0}(\mathbf{Z}, \mathbf{X}_{j_0i}) < \min_{1 \leq i \leq n_j} \rho_j(\mathbf{Z}, \mathbf{X}_{ji})$ for all $j \neq j_0$. Similarly, for the MCH classifier, one can compute

$$\rho_j^*(\mathbf{Z}, \mathbf{X}_{ji}) = \|\mathbf{Z} - \mathbf{X}_{ji}\| - \frac{1}{2} \binom{n_j}{2}^{-1} \sum_{s < t} \|\mathbf{X}_{js} - \mathbf{X}_{jt}\| \text{ for } j = 1, 2, \dots, J, \ i = 1, 2, \dots, n_j$$

and assign \mathbf{Z} to class j_0 if $\min_{1 \leq i \leq n_{j_0}} \rho_{j_0}^*(\mathbf{Z}, \mathbf{X}_{j_0i}) < \min_{1 \leq i \leq n_j} \rho_j^*(\mathbf{Z}, \mathbf{X}_{ji})$ for all $j \neq j_0$.

For the sake of simplicity, in Examples 1-3, we considered binary classification problems involving two normal distributions each having i.i.d. measurement variables. Now, one may be curious to know how CH and MCH classifiers perform in high-dimensional multi-class classification problems involving more general class distributions with possibly dependent and non-identically distributed measurement variables. For this investigation, we consider the following assumptions.

- (A1) In each of the J competing classes, the measurement variables have uniformly bounded fourth moments.
- (A2) If $\mathbf{X} = (X_1, \dots, X_d)^\top \sim F_j$ and $\mathbf{Y} = (Y_1, \dots, Y_d)^\top \sim F_i$ ($1 \leq j, i \leq J$) are independent, for $\mathbf{U} = \mathbf{X} - \mathbf{Y}$, $\sum_{r \neq s} |Corr(U_r^2, U_s^2)|$ is of the order $o(d^2)$.
- (A3) Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ be the mean vector and the dispersion matrix of F_j ($1 \leq j \leq J$). For each $j = 1, \dots, J$, there exists a constant σ_j^2 such that $\text{trace}(\boldsymbol{\Sigma}_j)/d \rightarrow \sigma_j^2$ as $d \rightarrow \infty$. Also, for each $i \neq j$, there exists a constant ν_{ji}^2 such that $\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|^2/d \rightarrow \nu_{ji}^2$ as $d \rightarrow \infty$.

Under (A1) and (A2), we have the weak law of large numbers (WLLN) (see, *e.g.*, Feller, 1991) for the sequence of possibly dependent and non-identically distributed random variables $\{U_q^2 : q \geq 1\}$, *i.e.*, $\left| \frac{1}{d} \|\mathbf{U}\|^2 - E\left(\frac{1}{d} \|\mathbf{U}\|^2\right) \right| \xrightarrow{P} 0$ as $d \rightarrow \infty$ (note that if the measurement variables are i.i.d., as they were in Examples 1-3, the WLLN holds under the second moment assumption, (A1) and (A2) are not needed there). Assumption (A3) gives the limiting value of $E\left(\frac{1}{d} \|\mathbf{U}\|^2\right)$ and hence that of $\frac{1}{d} \|\mathbf{U}\|^2 = \frac{1}{d} \|\mathbf{X} - \mathbf{Y}\|^2$ for $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ ($1 \leq j, i \leq J$). So, under (A1)-(A3), we have high-dimensional convergence of all pairwise distances and their limiting values (see Lemma 1 in Appendix). These assumptions are quite standard in the HDLSS literature. Hall *et al.* (2005) considered the d -dimensional observations as time series truncated at time d , and in addition to (A1) and (A3), they assumed the ρ -mixing property of the time series to study the high dimensional behavior of some popular classifier as d increases. Note that (A2) holds under that ρ -mixing condition. François *et al.* (2007) observed that for high-dimensional data with highly correlated or dependent measurement variables, pairwise distances are less concentrated than if all variables are independent. They claimed that the distance concentration phenomenon depends on the intrinsic dimension (see, *e.g.*, Levina and Bickel, 2004; Camastra and Staiano, 2016) of the data, instead of the dimension of the embedding space. So, in order to have distance concentration in high dimensions, one needs high intrinsic dimensionality of the data or weak dependence among the measurement variables. The assumption (A2) ensures that weak dependence. Some other similar relevant conditions for the convergence of pairwise distances can be found in (Ahn *et al.*, 2007; Jung and Marron, 2009; Sarkar and Ghosh, 2019; Yata and Aoshima, 2020; Banerjee and Ghosh, 2025). Under (A1)-(A3), we have the following theorem on the misclassification rates of the usual nearest neighbor, CH and MCH classifiers.

Theorem 1: If J competing classes satisfy assumptions (A1)-(A3), and there are at least two observations from each of them (*i.e.*, $n_j \geq 2$ for all $j = 1, 2, \dots, J$), then we have the following results.

- (a) If $\nu_{ji}^2 > |\sigma_j^2 - \sigma_i^2|$ for all $j \neq i$, the misclassification probability of the k -nearest neighbor classifier with $k < \min\{n_1, \dots, n_J\}$ converges to 0 as the dimension d grows to infinity. However, if $\nu_{ji}^2 < |\sigma_j^2 - \sigma_i^2|$ for some $j \neq i$, all observations from at least one class is misclassified with probability tending to 1 as d diverges to infinity.
- (b) If $\nu_{ji}^2 > 0$ for $j \neq i$, the misclassification probability of the CH classifier converges to 0 as d grows to infinity.
- (c) Suppose that for all $j \neq i$, either $\nu_{ji}^2 > 0$ or $\sigma_j^2 \neq \sigma_i^2$. Then, the misclassification probability of the MCH classifier converges to 0 as d grows to infinity.

Note that in Examples 1-3, we had $\nu_{12}^2 = \mu^2$, $\sigma_1^2 = 1$ and $\sigma_2^2 = \sigma^2$. The condition $\nu_{12}^2 > |\sigma_1^2 - \sigma_2^2|$ was violated in Examples 2 and 3, whereas the condition $\nu_{12}^2 > 0$ was also violated in Example 3. We had poor performance of NN and CH classifiers in these respective cases. But the condition $\nu_{12}^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$ was satisfied in all three examples. Consequently, the MCH classifier had good high-dimensional performance.

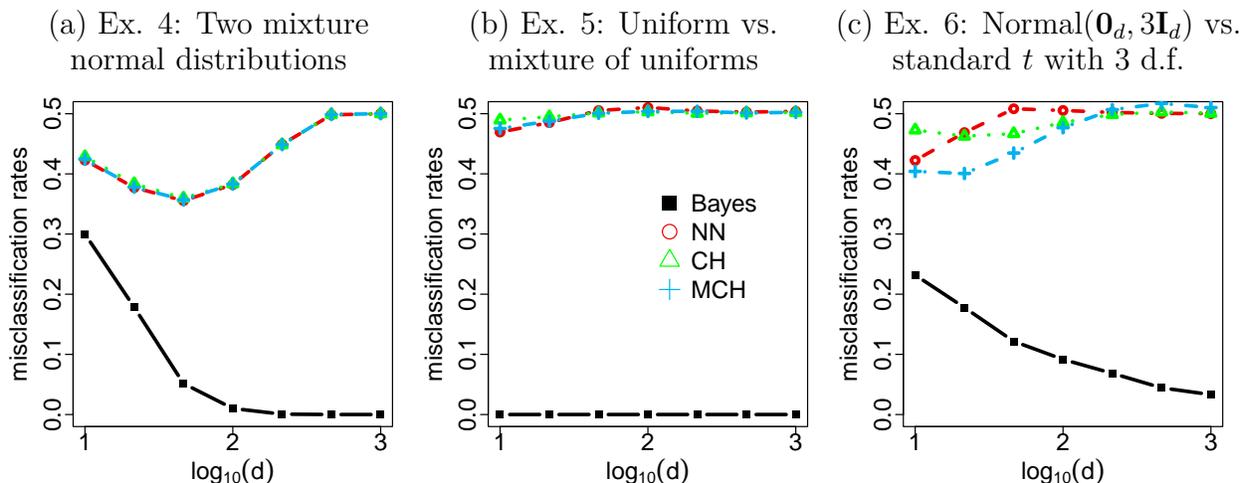


Figure 2: Misclassification rates of Bayes, NN, CH, and MCH classifiers in Examples 4-6.

Now, we consider three more examples (Examples 4-6) to investigate how the MCH classifier performs when at least one of the assumptions of Theorem 1 is violated.

Example 4: Each of the two classes is an equal mixture of two normal distributions. While one class is a mixture of $N(\mathbf{0}_d, \mathbf{I}_d)$ and $N(\mathbf{1}_d, 2\mathbf{I}_d)$, the other one is a mixture of $N(\boldsymbol{\alpha}_d, \mathbf{I}_d)$ and $N((\mathbf{1}_d - \boldsymbol{\alpha}_d), 2\mathbf{I}_d)$, where $\boldsymbol{\alpha}_d$ is a d -dimensional vector with entries 0 and 1 at even and odd places, respectively.

Example 5: Here the two classes are $U_d(1, 1.5)$ and an equal mixture of $U_d(0.5, 1)$ and $U_d(1.5, 2)$, where $U_d(a, b)$ denotes the d -dimensional uniform distribution over the region $\{\mathbf{x} \in \mathbb{R}^d : a \leq \|S^{1/2}\mathbf{x}\| \leq b\}$ for $S = 0.5\mathbf{I}_d + 0.5\mathbf{1}_d\mathbf{1}_d^\top$.

Examples 4 and 5 are dealing with mixture distributions. Here, (A1)-(A3) hold for each of the four sub-classes, but (A2) is violated for both competing classes. We also consider the following example:

Example 6: In this example, the two competing classes are $N(\mathbf{0}_d, 3\mathbf{I}_d)$ and the standard multivariate t -distribution with 3 degrees of freedom.

In Example 6, (A2) is violated for the t -distribution. Moreover, since the two classes have the same mean vector and the same dispersion matrix, we have $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. For each example, we consider different values of d ranging between 10 and 1000, and in each case, we form training and test samples of size 50 and 500, respectively, taking an equal number of observations from each class. Each experiment is repeated 100 times to compute the average test set misclassification rates of different classifiers, and they are reported in Figure 2. In these examples, NN, CH and MCH classifiers, all had poor performance, and they had misclassification rates close to 0.5 in high dimensions. These examples clearly

show the necessity to develop some new methods for high dimensional nearest neighbor classification, particularly for the examples involving mixture distributions. In the next section, we propose and discuss some methods for this purpose.

3. Nearest neighbor classification using distance-based features

In the previous sections, we have seen that for high dimensional classification based on nearest neighbors, the scale adjustment methods (CH and MCH) may not always be helpful. To take care of this problem, we suggest extracting some distance-based features from the data and constructing a suitable classifier on that feature space.

3.1. Classification based on minimum distances

Note that in a binary classification problem with training samples $\{\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n_1}\}$ and $\{\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2n_2}\}$ from the two competing classes, for classification of a test case \mathbf{Z} , the 1-NN classifier computes its minimum distances $d_1(\mathbf{Z}) = \min_{1 \leq i \leq n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|$ and $d_2(\mathbf{Z}) = \min_{1 \leq i \leq n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|$ from Class-1 and Class-2, respectively. Then it classifies \mathbf{Z} to the first class if $d_2(\mathbf{Z}) > d_1(\mathbf{Z})$ or $d_2^2(\mathbf{Z}) > d_1^2(\mathbf{Z})$. Like the 1-NN classifier, the CH classifier

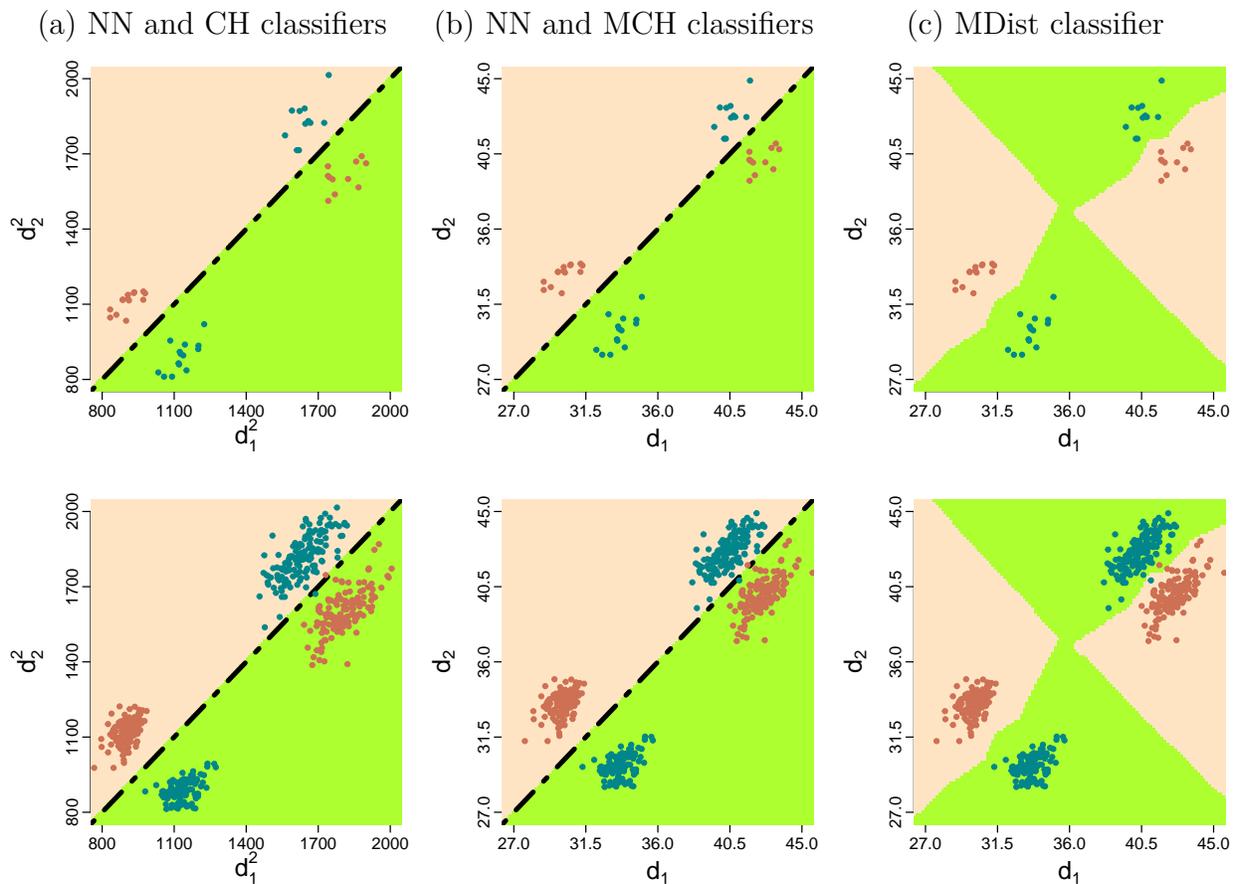


Figure 3: Scatter plots of training (top row) and test (bottom row) samples along with the class boundaries estimated by NN, CH, MCH, and MDist classifiers in Example 4.

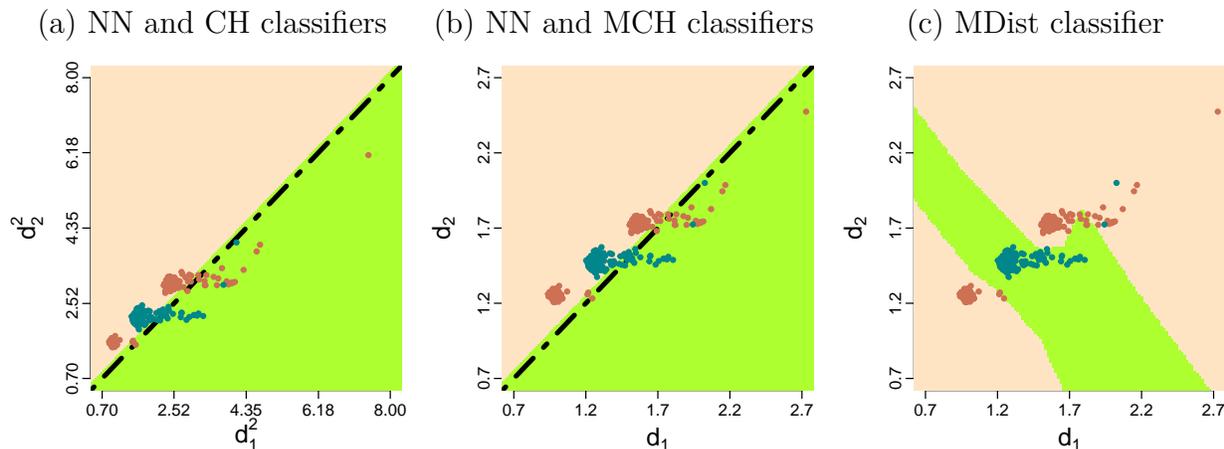


Figure 4: Scatter plots of the test samples and the class boundaries estimated by NN, CH, MCH and MDist classifiers in Example 5.

also leads to linear classification in the $d_1^2 - d_2^2$ space and classifies \mathbf{Z} to the first class if $d_2^2(\mathbf{Z}) > d_1^2(\mathbf{Z}) + C_1$, where $C_1 = \frac{1}{2} \left[\binom{n_2}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{2s} - \mathbf{X}_{2t}\|^2 - \binom{n_1}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{1s} - \mathbf{X}_{1t}\|^2 \right]$. Similarly, the MCH classifier leads to linear classification in the $d_1 - d_2$ space and classifies \mathbf{Z} to the first class if $d_2(\mathbf{Z}) > d_1(\mathbf{Z}) + C_2$, where $C_2 = \frac{1}{2} \left[\binom{n_2}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{2s} - \mathbf{X}_{2t}\| - \binom{n_1}{2}^{-1} \sum_{s<t} \|\mathbf{X}_{1s} - \mathbf{X}_{1t}\| \right]$. The first and the second columns in Figure 3 show the class boundaries estimated by these classifiers (the back line in the first and the second column shows the class boundary estimated by the 1-NN classifier) in Example 4 for dimension 500. They also show the scatter plots of $(d_1(\cdot), d_2(\cdot))$ (or $(d_1^2(\cdot), d_2^2(\cdot))$) for all training (top row) and test (bottom row) sample observations. For the training data points, the leave-one-out method (see, *e.g.*, Wong, 2015) is used to compute its minimum distances from the two classes. From this figure, it is quite evident that minimum distances (or squared minimum distances) from the two classes contain substantial information about class separability, but the resulting data clouds from the two classes are not linearly separable in that space. As a result, NN, CH, and MCH classifiers, all had poor performance. But we can overcome this problem if we use a suitable nonlinear classifier in that space. For instance one can use the 1-NN classifier in the $d_1 - d_2$ space. This classifier, which is referred to as the MDist classifier, performed well in this example. The last column in Figure 3 shows the class boundary estimated by the MDist classifier. Note that it correctly classified almost all observations.

We observed a similar phenomenon in Example 5 as well (see Figure 4). Like Example 4, here also the observations from different sub-classes form distinct clusters in the $d_1 - d_2$ space (or the $d_1^2 - d_2^2$ space). So, this feature space contains useful information about class separability, but the feature vectors from the two classes are not linearly separable. Therefore, while NN, CH and MCH classifiers had misclassification rates close to 50%, the MDist classifier had an excellent performance.

In Example 6, we have the convergence of pairwise distances for observations from the normal distribution, but not for observations from the multivariate t distribution. In this example, NN, CH and MCH classifiers classified almost all observations into a single class (see Figure 5), but the MDist classifier had much superior performance.

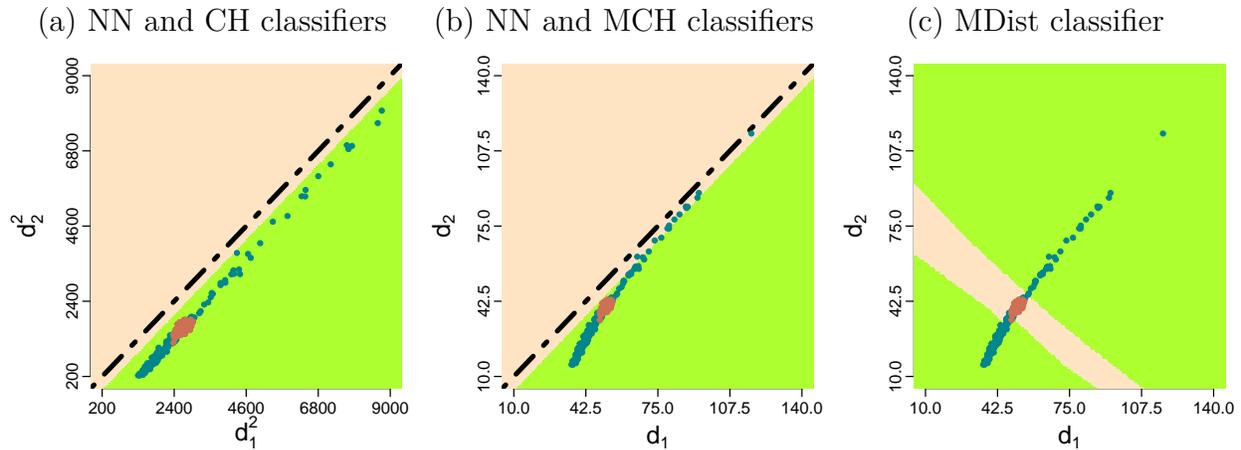


Figure 5: Scatter plots of the test samples and the class boundaries estimated by NN, CH, MCH and MDist classifiers in Example 6.

A similar idea of projecting the observations into a distance-based feature space and using the nearest neighbor classifier on that space was also considered in Dutta and Ghosh (2016), where the authors suggested using average distances $\bar{d}_1(\mathbf{Z}) = \text{avg}_i \|\mathbf{Z} - \mathbf{X}_{1i}\|$ and $\bar{d}_2(\mathbf{Z}) = \text{avg}_i \|\mathbf{Z} - \mathbf{X}_{2i}\|$ from the competing classes as features. Figure 6 shows these features for the test sample observations in Examples 4-6 and also the class boundaries estimated by the resulting classifier, called the TRAD classifier (see Dutta and Ghosh, 2016). Here also, for computing the feature vectors for the training sample observations, the leave-one-out method is used. Figure 6(a) shows that in Example 4, we have reasonable separability in the feature space, but the four distinct clusters are not as prominent as they were in Figure 3. Here we have some overlaps between the clusters corresponding to two competing classes. As a result, TRAD performed better than NN, CH and MCH classifiers, but not as good as the MDist classifier. This is also evident from Figure 7(a), which shows the average (over 100 replications) test set misclassification rates of these classifiers for various choices of d . In

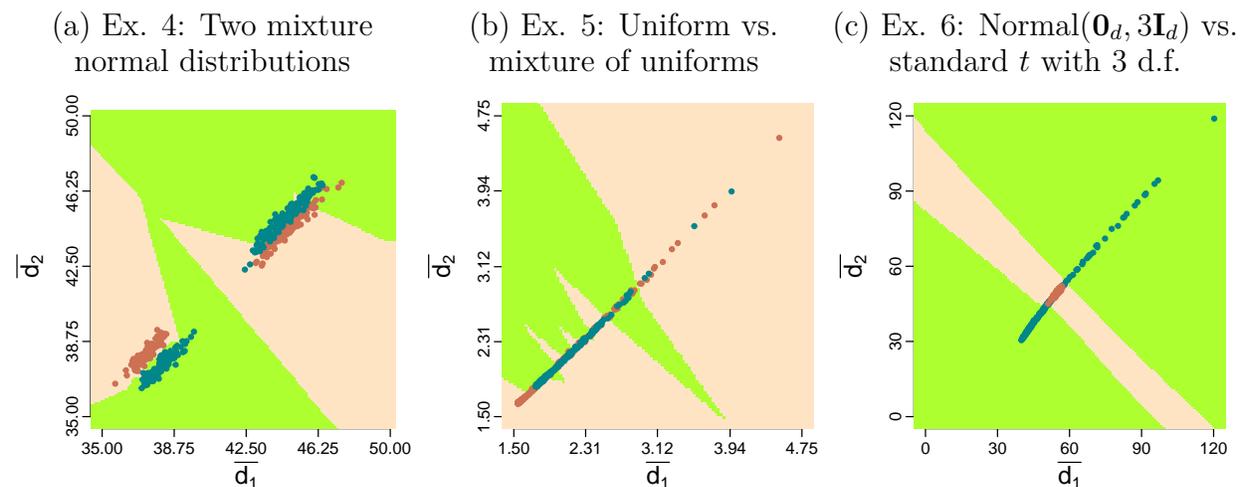


Figure 6: Scatter plots of the test samples and the class boundaries estimated by the TRAD classifier in Examples 4-6.

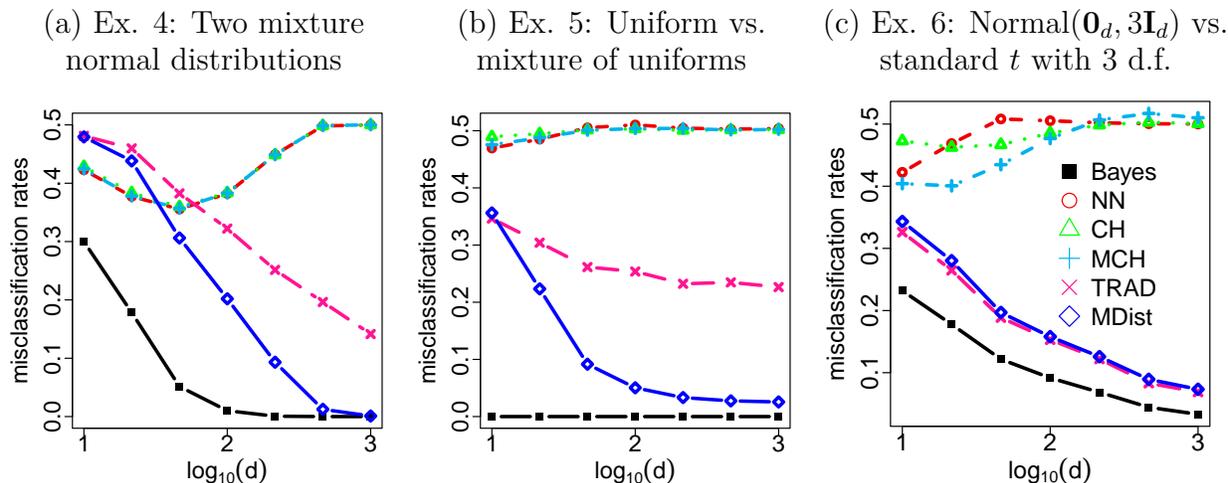


Figure 7: Misclassification rates of Bayes, CH, MCH, TRAD and MDist classifiers in Examples 4-6.

Example 5, the features based on average distances do not provide much separation between the two classes (see Figure 6(b)). So, as expected, TRAD had much higher misclassification rates (see 7(b)). However, in Example 6, the $\bar{d}_1 - \bar{d}_2$ space shows almost the same degree of separation as in the $d_1 - d_2$ space (see Figure 6). So, the class boundaries estimated by TRAD and MDist classifiers were almost similar, and they had almost similar misclassification rates. (see Figures 7(c)).

The success of the MDist classifier motivates us to carry out some theoretical analysis to understand its high-dimensional behavior. For this investigation, we again consider assumptions (A1)-(A3), and prove the perfect classification property of the MDist classifier in high dimensions.

Theorem 2: Suppose that J competing classes satisfy assumptions (A1)-(A3), and from each of them, there are at least two observations (*i.e.*, $n_j \geq 2$ for all $j = 1, 2, \dots, J$). If for all $j \neq i$, $\nu_{ji}^2 > 0$ or $\sigma_j^2 \neq \sigma_i^2$, the misclassification rate of the MDist classifier converges to 0 as d grows to infinity.

However, as we have discussed before, if the competing classes are mixtures of several sub-classes, the assumptions (A1)-(A3) may hold for each of the sub-classes but none of the competing classes (as in Examples 4). We have seen that in such situations, CH and MCH classifiers often have poor performance in high dimensions. However, from the proof of Theorem 2 (see Appendix), it is clear that in such cases, for each of the sub-classes, the feature vectors of minimum distances converge to a point as the dimension increases. If these points are distinct for each sub-class, we get some distinct clusters in the feature space, and the MDist classifier leads to perfect classification. We have already seen that in Example 4. A theorem similar to Theorem 2 can be stated for these mixture distributions as well, but the conditions for perfect classification by the MDist classifier (*i.e.*, the conditions needed to ensure that for any two sub-classes from two competing classes, the feature vectors converge to two distinct points, one for each sub-class) becomes mathematically complicated to interpret. That is why we choose not to state that theorem here.

Now we consider two interesting examples (Example 7 and 8) involving binary classification, where each of the two competing classes satisfies assumptions (A1)-(A3), but we have $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. So, in this case, the feature vectors $(d_1(\cdot), d_2(\cdot))$ corresponding to two competing classes converge to the same limiting value as d increases. One may be curious to know how the MDist classifier performs in such situations, and we investigate it here. Here also, we consider different values of d ranging between 10 and 1000, form the training and the test sets of size 50 and 500 by taking an equal number of observations from the two classes and repeat the experiment 100 times to compute the average test set misclassification rates of different classifiers.

Example 7: We consider two normal distributions having the same mean vector $\mathbf{0}_d$ but different dispersion matrices $\Lambda_1 = \text{diag}(\lambda_{11}, \dots, \lambda_{1d})$ and $\Lambda_2 = \text{diag}(\lambda_{21}, \dots, \lambda_{2d})$. Here $\lambda_{1i} = 1/2$ and $\lambda_{2i} = 2$ for $i \leq d/2$, whereas $\lambda_{1i} = 2$ and $\lambda_{2i} = 1/2$ for $i > d/2$.

Figure 8(a) show the scatter plots of the test set observations in the $d_1 - d_2$ space and the class boundary estimated by the MDist classifier for $d = 500$. It is clear that unlike previous examples, here the features based on minimum distances fail to discriminate between the two classes. As a result, the MDist classifier had much higher misclas-

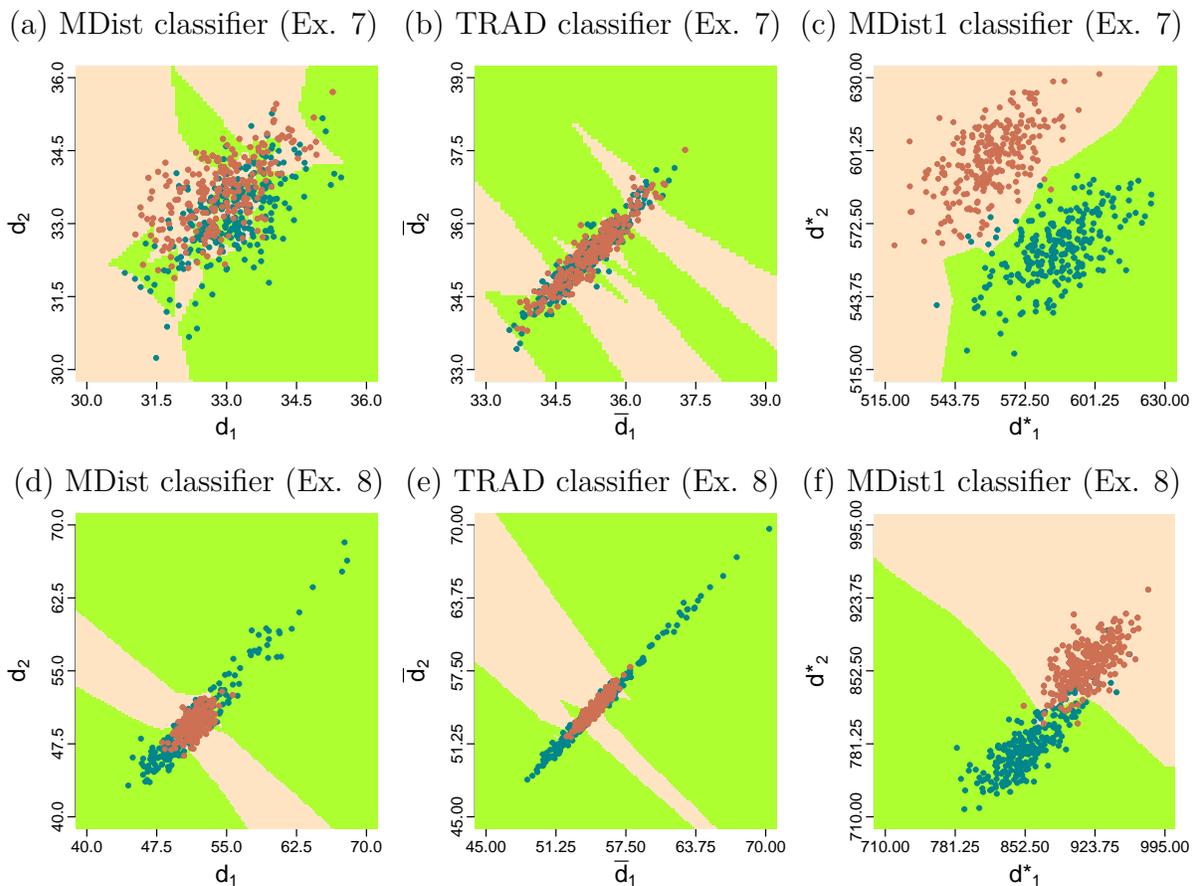


Figure 8: Scatter plots of the test samples and the class boundaries estimated by the MDist, TRAD and MDist1 classifiers in Example 7 (top row) and Example 8 (bottom row).

(a) Ex. 7: Two normals with different dispersion matrices having same trace

(b) Ex. 8: Product of $N(0, 3)$ vs product of univariate t_3

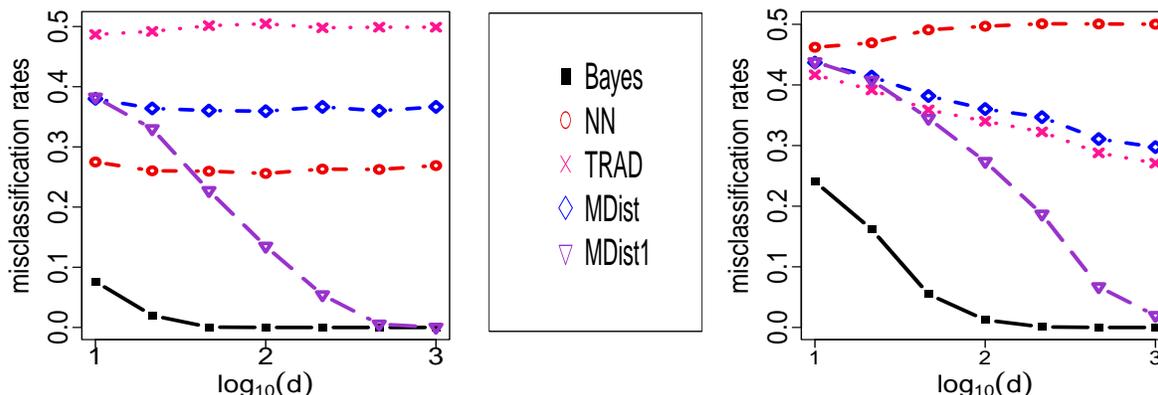


Figure 9: Misclassification rates of Bayes, NN, TRAD, MDist and MDist1 classifiers in Examples 7 and 8.

sification rates (see Figure 9(a)). The features based on average ℓ_2 distances also fail to provide any discriminatory information (see Figure 8(b)) in this case. The performance of the TRAD classifier was even worse. It misclassified almost half of the test set observations (see Figure 9(a)). Surprisingly, in this example, we get a good result if instead of ℓ_2 distance (Euclidean distance), we use ℓ_1 distance (Manhattan distance) for finding the neighbors. If $\{\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n_1}\}$ and $\{\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2n_2}\}$ are training sample observations from two competing classes (here we have $n_1 = n_2 = 25$), for any \mathbf{Z} , we can use $d_1^*(\mathbf{Z}) = \min_{1 \leq i \leq n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|_1$ and $d_2^*(\mathbf{Z}) = \min_{1 \leq i \leq n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|_1$ as features and perform usual nearest neighbor classification on that feature space. Here also for computing d_1^* and d_2^* at the training data points, we use the leave-one-out method. Figure 8(c) shows the scatter plot of these features for all test set observations and the class boundary estimated by the 1-NN classifier on this feature space (we call it the MDist1 classifier). Here we have two distinct clusters in the feature space, one for each class. As a result, the MDist1 classifier had an excellent performance and correctly classified almost all observations. The average test set misclassification rate of this classifier (reported in Figure 9(a)) also tells us the same story. Now, let us consider the following example.

Example 8: Here each of the two classes has *i.i.d.* measurement variables. In Class-1, they follow the $N(0, 3)$ distribution, while in Class-2, they follow the standard t distribution with 3 degrees of freedom.

Note that this is different from the multivariate t distribution considered in Example 6. In this example also, the features based on minimum ℓ_2 distances and those based on average ℓ_2 distances do not provide much separability between the two classes (see Figure 8(d) and (e)), but the features based on minimum ℓ_1 distances make the data clouds better separated (see Figure 8(f)). As a result, the MDist1 classifier outperformed TRAD and MDist classifiers (see Figure 9(b)).

To understand the high-dimensional behavior of the MDist1 classifier, we carry out some theoretical investigations under the following assumptions, which are similar to (A1)-(A3) stated before.

- (A1^o) In each of the J competing classes, the measurement variables have uniformly bounded second moments.
- (A2^o) If $\mathbf{X} = (X_1, \dots, X_d)^\top \sim F_j$ and $\mathbf{Y} = (Y_1, \dots, Y_d)^\top \sim F_i$ ($1 \leq j, i \leq J$) are independent, for $\mathbf{U} = \mathbf{X} - \mathbf{Y}$, $\sum_{r \neq s} |Corr(|U_r|, |U_s|)|$ is of the order $o(d^2)$.
- (A3^o) For independent random vectors $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ ($1 \leq j, i \leq J$), $E\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|_1\right) = \frac{1}{d} \sum_{q=1}^d E|X_q - Y_q|$ converges to a constant τ_{ji} as $d \rightarrow \infty$.

Under (A1^o) and (A2^o), we have the convergence of pairwise ℓ_1 distances. Following similar steps as used in the proof of Lemma 1, one can show that for $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ ($1 \leq j, i \leq J$), $\left|\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|_1 - E\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|_1\right)\right| \xrightarrow{P} 0$ as $d \rightarrow \infty$.

For any $q = 1, 2, \dots, d$, define $e_{ji}^{(q)} = 2E|X_q - Y_q| - E|X_q - X'_q| - E|Y_q - Y'_q|$, where $\mathbf{X}, \mathbf{X}' \sim F_j$ and $\mathbf{Y}, \mathbf{Y}' \sim F_i$ ($j \neq i$) are independent random vectors. This quantity $e_{ji}^{(q)}$ can be viewed as the energy distance (see, *e.g.*, Székely and Rizzo, 2023) between the q -th marginals of F_j and F_i ($F_j^{(q)}$ and $F_i^{(q)}$, say). Following Baringhaus and Franz (2004), one can show that $e_{ji}^{(q)}$ is non-negative, and it takes the value 0 if and only if $F_j^{(q)} = F_i^{(q)}$. So, for any fixed dimension d , we have $\frac{1}{d}[2E\|\mathbf{X} - \mathbf{Y}\|_1 - E\|\mathbf{X} - \mathbf{X}'\|_1 - E\|\mathbf{Y} - \mathbf{Y}'\|_1] = \frac{1}{d} \sum_{q=1}^d e_{ji}^{(q)} = \bar{e}_{ji}(d) \geq 0$, where the equality holds if and only if $F_j^{(q)} = F_i^{(q)}$ for $q = 1, 2, \dots, d$. Therefore, it is somewhat reasonable to assume that $\mathcal{E}_{ji} = \lim_{d \rightarrow \infty} \bar{e}_{ji}(d) > 0$, which essentially says that the average coordinate-wise energy distance is asymptotically non-negligible. Under this assumption, we have the perfect separation property of the MDist1 classifier, which is asserted by the following theorem.

Theorem 3: Suppose that J competing classes satisfy assumptions (A1^o)-(A3^o), and from each of them, there are at least two observations (*i.e.*, $n_j \geq 2$ for all $j = 1, 2, \dots, J$). If the limiting value of the average coordinate-wise energy distance $\mathcal{E}_{ji} > 0$ for all $j \neq i$, the misclassification rate of the MDist1 classifier converges to 0 as d grows to infinity.

Note that while TRAD and MDist classifiers fail to discriminate between two distributions differing outside the first two moments, the MDist1 classifier can discriminate between them as long as they differ in their one-dimensional marginals. That is why it outperformed TRAD and MDist classifiers in Examples 7 and 8.

This classifier enjoys the perfect separation property in high dimensions even when the competing classes are mixtures of several sub-classes, and these sub-class distributions satisfy assumptions (A1^o)-(A3^o). It becomes clear from the proof of Theorem 3 (see Appendix) that in such cases, for each of the sub-classes, the feature vectors of minimum ℓ_1 distances (after appropriate scaling) converge to a point as the dimension increases. If these points are distinct for each sub-class, the MDist1 classifier leads to perfect classification. But here also writing the conditions for perfect classification becomes mathematically complicated to interpret. So, we decide not to state another theorem in this regard. However, our analysis of simulated data sets clearly demonstrates this. Figure 10 shows the misclassification rates of TRAD, MDist and MDist1 classifiers in Examples 1-6 along with those of 1-NN and Bayes

classifiers. In all these examples including Example 4 and 5, where we deal with mixture distributions, MDist and MDist1 classifiers had similar performance.

This figure shows another interesting phenomenon. In Examples 1-3, when the underlying distributions are unimodal, the TRAD classifier, which considers the average of distances from all observations, performed better than MDist and MDist1 classifiers, which consider the distance of one nearest neighbor only. In Example 6 also, TRAD had an edge over the other two classifiers. But, in the case of mixture distributions (see Examples 4 and 5), taking the average over all observations coming from different sub-classes does not seem to be a meaningful option. In those cases, MDist and MDist1 classifiers outperformed the TRAD classifier. These result shows that instead of always going for features based on a single nearest neighbor from each class, sometimes it is more meaningful to consider distances from multiple nearest neighbors. We can include these distances in the set of features and go for classification in the extended feature space. We consider such methods in the following subsection.

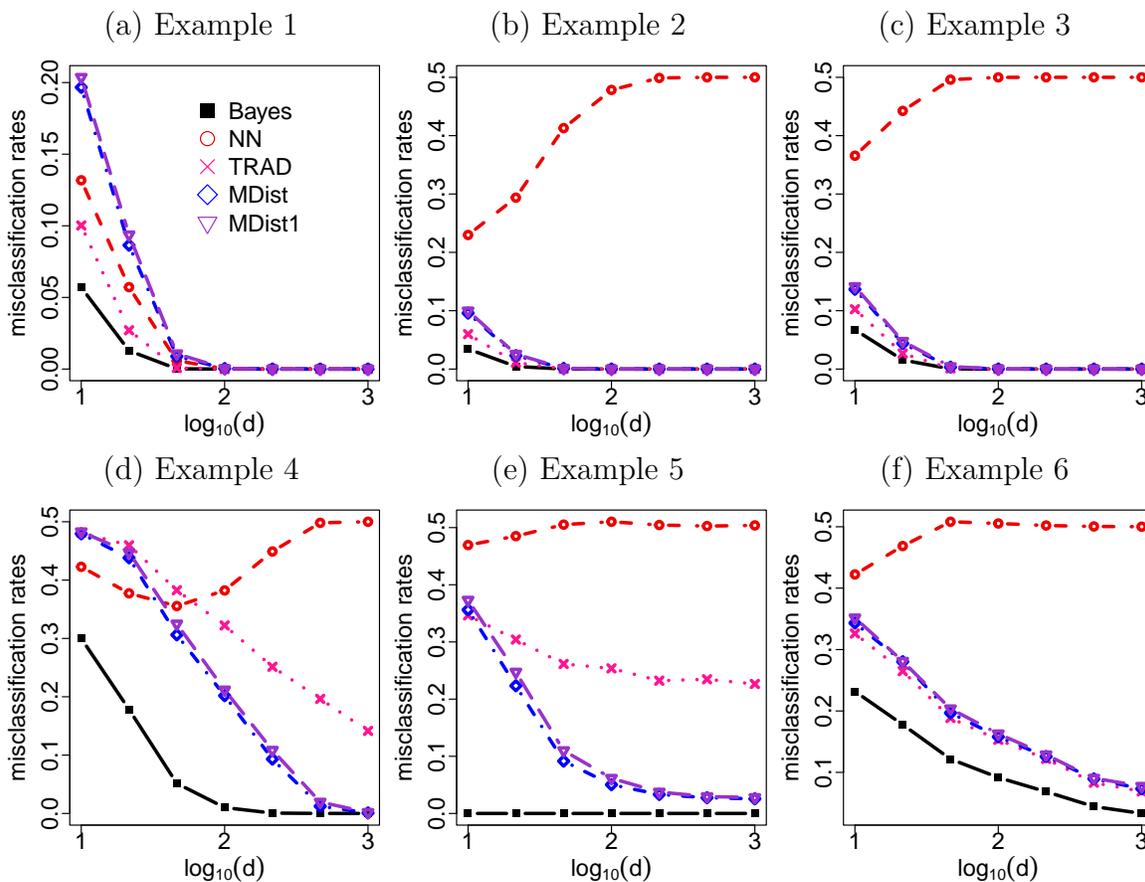


Figure 10: Misclassification rates of Bayes, NN, TRAD, MDist and MDist1 classifiers in Examples 1-6.

3.2. Classification based on multiple neighbors

Instead of considering only the distance of the first neighbor from each class, here we consider the distances of the first r ($r \geq 1$) neighbors from each class and use them

as features. So, if there are J competing classes, we consider a total of Jr many features and use the 1-NN classifier on that feature space. Here also, we can use ℓ_2 distances or ℓ_1 distances as features, and the resulting classifier is referred to as rMDist and rMDist1 classifiers, respectively. One may also consider both ℓ_1 and ℓ_2 distances and deal with $2Jr$ many features simultaneously. We refer to the resulting classifier as the rMDistC classifier. In all these cases the value of r is chosen by minimizing the leave-one-out cross-validation

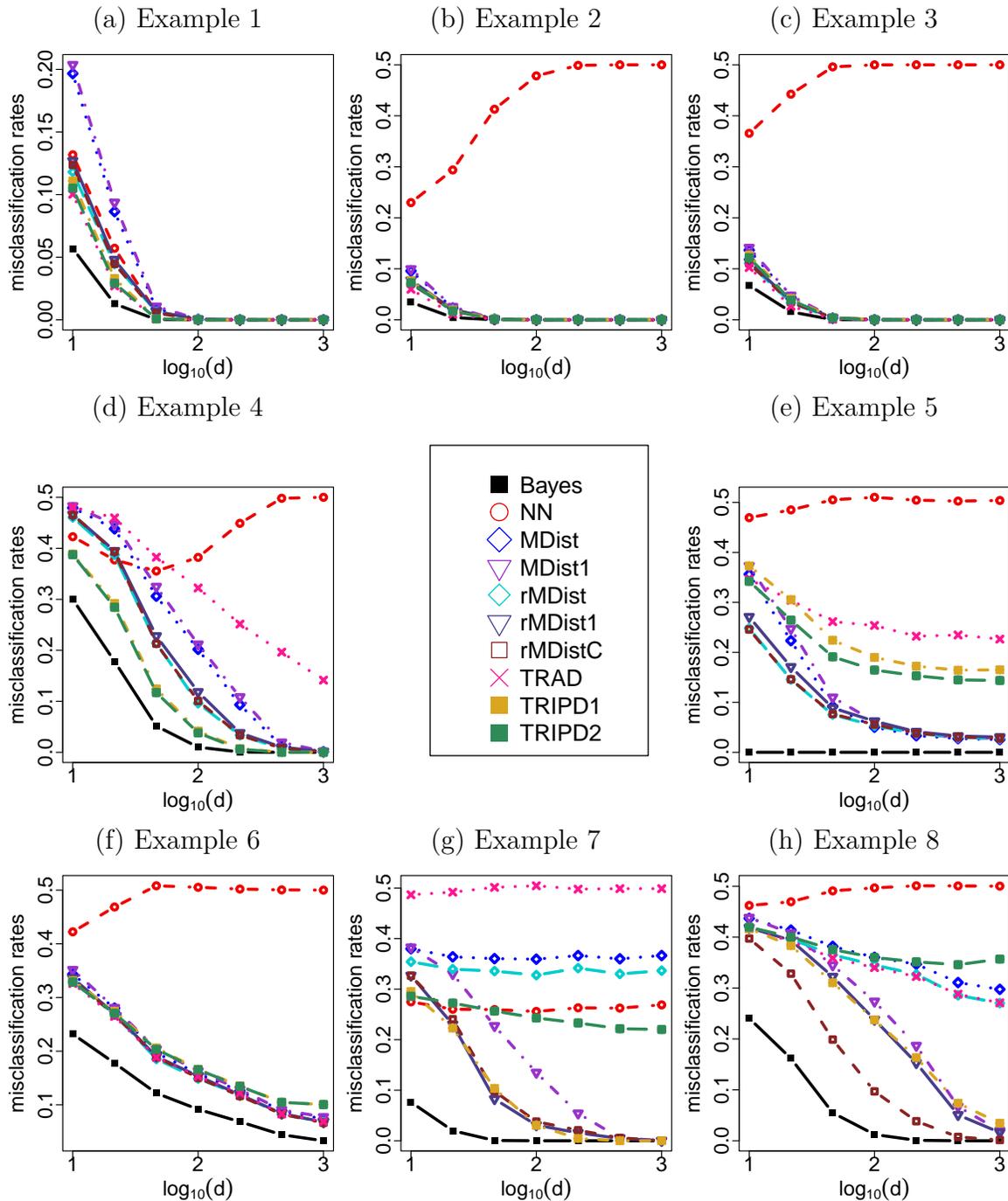


Figure 11: Misclassification rates of different classifiers in Examples 1-8.

estimate (see, *e.g.* Wong, 2015) of the misclassification rate. Figure 11 shows the average test set misclassification rates of these classifiers in Examples 1-8. One can see that in most of the examples, rMDist and rMDist1 classifiers performed better than MDist and MDist1 classifiers, respectively. The rMDistC classifier also performed well in almost all examples. In Examples 7 and 8, it outperformed rMDist and rMDist1 classifiers.

A somewhat similar classification method was considered in Dutta and Ghosh (2016), where the authors transformed each observation into an n -dimensional vector containing its distances from all training sample observations. They also considered ℓ_1 and ℓ_2 distances for transformation and called the resulting classifiers as TRIPD1 and TRIPD2. respectively. Misclassification rates of those two methods are also reported in Figure 11. In Examples 1-3 and 6, their misclassification rates were similar to our proposed methods. In Example 4, they had the lowest misclassification rates, but in Example 5, they were outperformed by our proposed classifiers. In Examples 7 and 8, the TRIPD2 classifier, which is based on ℓ_2 distances, had higher misclassification rates. In Example 7, the performance of the TRIPD1 classifier was comparable to rMDist1 and rMDistC classifiers, but in Example 8, the rMDistC classifier had a clear edge.

4. Results from the analysis of benchmark datasets

We analyze 10 benchmark datasets for further evaluation of the performance of the proposed and existing methods discussed in the previous two sections. For these benchmark datasets, since the true class distributions are not known, it is not possible to compute the Bayes risks. Therefore, to facilitate comparison, we report the misclassification rates of two popular classifiers, support vector machines (SVM) (see, *e.g.* Cristianini and Shawe-Taylor, 2003; Scholkopf and Smola, 2018) and random forest (RF) (see, *e.g.* Breiman, 2001; Genuer and Poggi, 2020), which are known to perform well for high dimensional data. Since the nearest neighbor classifiers are nonlinear, to make it fair, here we use the nonlinear SVM for comparison. For our numerical study, we use the radial basis function kernel, where all tuning parameters are chosen using the 5-fold cross-validation method (see, *e.g.*, Wong, 2015). We use the R package `caret` for this purpose. The same package is used for the random forest classifier as well, where we use default tuning parameters.

Out of these 10 datasets, Chowdary and Nutt datasets are taken from CompCancer dataset-Schliep lab. The rest of the datasets are taken from the UCR Time Series Classification Archive. Detailed descriptions of these datasets are available at these respective sources. The datasets taken from the UCR archive have specific training and test sets. We merge these two sets and divide the pooled dataset randomly into two parts to form the training and the test samples. Except for the Synthetic Control Chart data, in all other cases, the sizes of training and test samples are taken to be the same as they are in the data archive. Note that in all these cases, the size of the training sample is smaller than the dimension. In the case of Synthetic Control Chart data, instead of an equal partition (as in UCR archive), we use training and test samples of size 60 and 540, respectively, so that the training sample size does not become larger than the dimension. The datasets from the CompCancer database do not have specific training and test samples. In these cases, we divide the data sets into equal halves to form the training and the test samples. Brief descriptions of these datasets are given in Table 1. In all these cases, we form the training and the test samples in such a way that the proportions of different classes in the two samples are

Table 1: Brief descriptions of the benchmark datasets.

dataset	d	J	Sample size		dataset	d	J	Sample size	
			Train	Test				Train	Test
Synthetic Control	60	6	60	540	Lightning7	319	7	70	73
Chowdary	182	2	52	52	Herring	512	2	64	64
Trace	275	4	100	100	Nutt	1070	2	14	14
Toe Segmentation1	277	2	40	228	Gordon	1628	2	90	91
Coffee	286	2	28	28	Colon Cancer	2000	2	31	31

Table 2: Average misclassification rates (in %) of different classifiers and their standard errors (reported inside the bracket) in benchmark datasets.

Dataset	Synth. Control	Chowdary	Trace	Toe Seg-ment.1	Coffee	Lightning7	Herring	Nutt	Gordon	Colon Cancer
NN	18.78 (0.28)	4.83 (0.29)	20.33 (0.37)	38.62 (0.36)	2.00 (0.31)	37.97 (0.43)	51.33 (0.59)	34.00 (0.92)	2.96 (0.16)	26.10 (0.65)
MDist	10.02 (0.26)	7.98 (0.45)	13.51 (0.46)	42.80 (0.45)	2.61 (0.34)	39.29 (0.47)	45.53 (0.52)	14.14 (0.74)	1.92 (0.17)	32.42 (0.95)
MDist1	12.29 (0.30)	7.04 (0.46)	18.88 (0.46)	37.30 (0.45)	4.43 (0.39)	35.72 (0.52)	47.20 (0.57)	15.71 (0.70)	1.19 (0.10)	34.74 (0.94)
rMDist	10.06 (0.28)	6.56 (0.37)	14.90 (0.48)	40.92 (0.38)	2.93 (0.32)	38.16 (0.42)	46.86 (0.54)	15.14 (0.86)	1.73 (0.14)	22.06 (0.92)
rMDist1	10.12 (0.27)	5.35 (0.35)	19.61 (0.44)	35.64 (0.47)	4.50 (0.39)	36.00 (0.44)	46.06 (0.63)	15.93 (0.68)	1.27 (0.10)	27.03 (1.03)
rMDistC	9.25 (0.27)	5.90 (0.32)	15.01 (0.49)	35.61 (0.47)	3.07 (0.34)	34.67 (0.47)	46.88 (0.54)	14.21 (0.80)	1.64 (0.15)	24.65 (0.96)
TRAD	14.78 (0.30)	7.31 (0.34)	24.48 (0.37)	49.93 (0.41)	4.11 (0.43)	37.28 (0.39)	48.08 (0.56)	12.07 (0.72)	6.52 (0.19)	18.06 (0.72)
TRIPD1	7.67 (0.19)	4.42 (0.25)	23.25 (0.43)	33.92 (0.32)	6.14 (0.40)	31.47 (0.39)	47.36 (0.53)	17.57 (0.95)	1.21 (0.10)	25.77 (0.66)
TRIPD2	6.42 (0.19)	5.38 (0.30)	21.08 (0.39)	38.15 (0.34)	3.79 (0.39)	32.27 (0.40)	50.50 (0.56)	8.93 (0.86)	3.38 (0.20)	21.58 (0.62)
Random Forest	13.07 (0.24)	5.00 (0.26)	14.23 (0.48)	38.21 (0.41)	3.21 (0.45)	28.29 (0.47)	40.30 (0.47)	21.07 (1.09)	0.93 (0.11)	29.55 (0.64)
Nonlin. SVM	9.50 (0.32)	7.63 (0.50)	10.48 (0.36)	45.02 (0.37)	4.25 (0.46)	35.77 (0.47)	39.11 (0.41)	11.21 (0.75)	2.53 (0.22)	20.90 (0.75)

as close as possible. In each case, this partitioning is carried out 100 times, and the average test set misclassification rates of different classifiers are reported in Table 2 along with their corresponding standard errors. Overall performances of CH and MCH classifiers (especially, that of the former one) were much inferior compared to all other classifiers considered here. Therefore, we do not report them in this section.

Though the 1-NN classifier had the lowest misclassification rate in the Coffee dataset and the second lowest misclassification rate in the Chowdary dataset, in many cases, its performance was far from the best one (see Table 2). For instance, in Nutt and Synthetic Control Chart datasets, its misclassification rates were much higher compared to all other classifiers considered here. Furthermore, it had the highest misclassification rate in the Her-

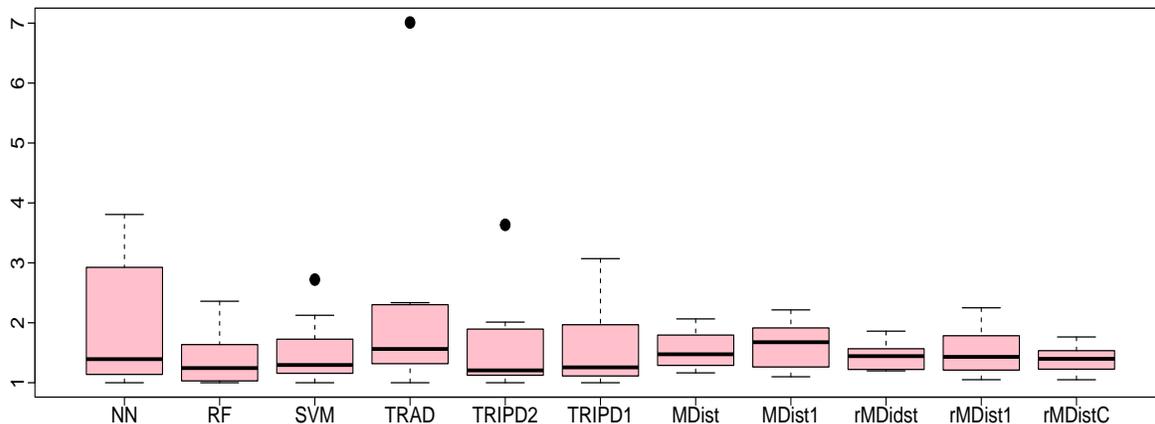


Figure 12: Boxplots showing the robustness of different classifiers in benchmark datasets.

ring data set. TRAD also had relatively higher misclassification rates in many examples (*e.g.*, Synthetic Control Chart, Trace, Toe Segmentation, and Gordon datasets). Only in the case of Colon Cancer data, it outperformed others. Our proposed classifiers had good performance in most of these examples. While the classifiers based on multiple neighbors (rMDist, rMDist1 and rMDistC) outperformed those based on single neighbors (MDist and MDist1) in Chowdary, Toe Segmentation, and Colon Cancer data sets, in all other cases, they had comparable performance. In Trace and Herring data sets, these classifiers performed better than TRAD, TRIPD1 and TRIPD2 classifiers. Table 2 clearly shows that the performances of our proposed classifiers, particularly for classifiers based on multiple neighbors, were comparable to nonlinear SVM and random forest.

To compare the overall performances of different classifiers concisely and comprehensively, we used the notion of robustness introduced in Friedman (1994). If there are T classifiers who have misclassification rates e_1, e_2, \dots, e_T in a particular data set, the robustness of the t -th classifier is computed as $R(t) = e_t/e_0$, where $e_0 = \min_{1 \leq t \leq T} e_t$. So, in an example, the best classifier has $R(t) = 1$, while higher values of $R(t)$ indicate the lack of robustness of the t -th classifier. For each of these benchmark data sets, we computed these ratios for all classifiers, and they are graphically represented using box plots in Figure 12. This figure clearly shows that the overall performances of all other classifiers were somewhat better than the usual nearest neighbor classifier. It also shows that among our proposed classifiers, those based on multiple neighbors performed better than the corresponding classifiers based on a single nearest neighbor. While the rMDist classifier exhibited better robustness properties than TRIPD2, the rMDist1 classifier turned out to be more robust than the TRIPD1 classifier. The rMDistC classifier, which considers both ℓ_1 and ℓ_2 distances of the nearest neighbors, also had an excellent overall performance. If not better, the performances of our proposed classifiers were comparable to the popular classifiers like nonlinear SVM and random forest.

5. Concluding remarks

In this article, we have proposed some possible modifications to the nearest neighbor classifier for the classification of high-dimensional data. We have seen that if the location

difference among the competing classes gets masked by their scale difference, the usual nearest neighbor classifier performs poorly in high dimensions. The adjustment proposed by Chan and Hall (2009b) takes care of this problem, but the resulting classifier fails when the competing classes differ outside their first moments. The MCH classifier overcomes this limitation, and it can discriminate between two high-dimensional distributions differing either in their locations or in their scales. However, this method may not work well in many situations, especially when the class distributions are mixtures of several widely varying subclasses. The proposed classifiers based on minimum distances are helpful in such situations. The MDist1 classifier can even discriminate among competing classes differing outside the first two moments. Instead of considering only one neighbor from each class, sometimes it is helpful to consider the distances of the first r neighbors and perform nearest neighbor classification in that feature space. Analyzing several simulated and benchmark datasets, we have amply demonstrated that if not better, our proposed classifiers yield competitive performance in high dimensions.

In this article, we have used nearest neighbor classification on the feature space of ℓ_1 or ℓ_2 distances. Though we have seen some theoretical advantages of using the ℓ_1 distance, our analysis of benchmark datasets clearly shows that in practice, there is no clear winner. So, a user may wonder which of the two feature spaces to use in a given problem. One may also like to use ℓ_p -distances for other choices of p or features based on other generalized distance functions of the form $\varphi_{h,\psi}(\mathbf{x}, \mathbf{y}) = h\left\{\frac{1}{d} \sum_{i=1}^d \psi(|x^{(i)} - y^{(i)}|^2)\right\}$ where $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ and $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are continuous, strictly increasing functions with $h(0) = \psi(0) = 0$ as introduced in Sarkar and Ghosh (2018). For suitable choices of h and ψ (e.g., when $h(t) = t$ and $\psi(t)$ has non-constant completely monotone derivatives), it ensures the positivity of the energy distance (as discussed in the paragraph before Theorem 3) in any finite dimension. Moreover, if ψ is bounded, it also makes the resulting classifier robust against outliers. However, the choice of the optimal features is a challenging problem, and it would be helpful if a data-driven method can be developed for this purpose. Throughout this article, for our proposed methods, we have always used the 1-NN classifier in the feature space. This is mainly for a fair comparison with other competing nearest neighbor methods (e.g., CH, TRAD, TRIPD1, and TRIPD2), where 1-NN classification is considered. However, in practice, one may use the k -NN classifier for other values of k as well. For constructing the rMDistC classifier, though we have considered the same number of ℓ_1 and ℓ_2 distances as features, it is possible to include r_1 many ℓ_1 distances and r_2 many ℓ_2 distance in the set of features. We avoid choosing different values of r_1 and r_2 to reduce the computing cost at the cross-validation step. In practice, distances from all of the first r neighbors may not always be important for classification. In such cases, a suitable feature selection criterion would be helpful. Instead of feature selection, one can also think about constructing an ensemble classifier (see, e.g. Dietterich, 2000; Zhang and Zhang, 2009; Kiziloz, 2021) like random forest, where we construct different classifiers based on different sets of features and judiciously aggregate them. These problems can be investigated in a separate work in future.

References

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

- Banerjee, B. and Ghosh, A. K. (2025). On high dimensional behaviour of some two-sample tests based on ball divergence. *Statistica Sinica*, **35**, To appear.
- Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**, 190–206.
- Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.
- Camastra, F. and Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, **328**, 26–41.
- Carrerira-Perpinan, M. (2009). A review of dimension reduction techniques. Technical report. Department of Computer Science, University of Sheffield.
- Chan, Y. B. and Hall, P. (2009a). Robust nearest neighbor methods for classifying high-dimensional data. *The Annals of Statistics*, **37**, 3186–3203.
- Chan, Y.-B. and Hall, P. (2009b). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, **96**, 469–478.
- Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21–27.
- Cristianini, N. and Shawe-Taylor, J. (2003). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.
- Deegalla, S. and Bostrom, H. (2006). Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *5th International Conference on Machine Learning and Applications*, pages 245–250. IEEE.
- Devroye, L., Györfi, L., and Lugosi, G. (2013). *A Probabilistic Theory of Pattern Recognition*. Springer Science & Business Media, Berlin.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2007). *Pattern Classification*. Wiley, New York.
- Dutta, S. and Ghosh, A. K. (2016). On some transformations of high dimension, low sample size data for nearest neighbor classification. *Machine Learning*, **102**, 57–83.
- Feller, W. (1991). *An Introduction to Probability Theory and its Applications, Volume 1*. John Wiley & Sons.
- Fern, X. Z. and Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 186–193.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.
- Fix, E. and Hodges, J. L. (1951). Discriminatory analysis - nonparametric discrimination: Consistency properties. Project 21-49-004, Report 4, US Air Force School of Aviation Medicine, Randolph Field.
- Fradkin, D. and Madigan, D. (2003). Experiments with random projections for machine learning. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 517–522.
- François, D., Wertz, V., and Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, **19**, 873–886.
- Friedman, J. H. (1994). Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University.

- Genuer, R. and Poggi, J.-M. (2020). *Random Forests*. Springer.
- 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**, 427–444.
- Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, **36**, 2135–2152.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, **37**, 4104–4130.
- Kiziloz, H. E. (2021). Classifier ensemble methods in feature selection. *Neurocomputing*, **419**, 97–107.
- Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, **17**.
- Macionczyk, R., Moryc, M., and Buchtyar, P. (2023). Analyzing the impact of principal component analysis on k-nearest neighbors and naive bayes classification algorithms. In *International Conference on Information and Software Technologies*, pages 247–263. Springer.
- Pal, A. K., Mondal, P. K., and Ghosh, A. K. (2016). High dimensional nearest neighbor classification based on mean absolute differences of inter-point distances. *Pattern Recognition Letters*, **74**, 1–8.
- Radovanovic, M., Nanopoulos, A., and Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, **11**, 2487–2531.
- Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B*, **10**, 159–203.
- Roy, S., Sarkar, S., Dutta, S., and Ghosh, A. K. (2022). On generalizations of some distance based classifiers for HDLSS data. *Journal of Machine Learning Research*, **23**, 1–41.
- Sarkar, S. and Ghosh, A. K. (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat*, **7**, e187, 16.
- Sarkar, S. and Ghosh, A. K. (2019). On perfect clustering of high dimension, low sample size data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2257–2272.
- Scholkopf, B. and Smola, A. J. (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.
- Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science and Business Media.
- Székely, G. J. and Rizzo, M. L. (2023). *The Energy of Data and Distance Correlation*. Chapman and Hall/CRC.
- Tomašev, N., Radovanović, M., Mladenović, D., and Ivanović, M. (2014). Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. *International Journal of Machine Learning and Cybernetics*, **5**, 445–458.
- Vempala, S. S. (2005). *The Random Projection Method*. DIMACS - Series in Discrete Mathematics and Theoretical Computer Science, Volume 65, American Mathematical Society.

- Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, **10**, 207–244.
- Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, **48**, 2839–2846.
- Yata, K. and Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, **47**, 899–921.
- Zhang, C.-X. and Zhang, J.-S. (2009). A novel method for constructing ensemble classifiers. *Statistics and Computing*, **19**, 317–327.

APPENDIX

Lemma 1: If J competing classes satisfy (A1)-(A3), for two independent random vectors $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ ($1 \leq j, i \leq J$), $\|\mathbf{X} - \mathbf{Y}\|^2/d \xrightarrow{P} \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$, where $\nu_{ji}^2 = 0$ for $j = i$.

Proof: Note that using Chebyshev’s inequality, for any $\epsilon > 0$, we get

$$P\left(\left|\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2 - E\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2\right)\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \text{Var}\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2\right).$$

Now, $\text{Var}\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2\right) = \frac{1}{d^2} \left[\sum_{s=1}^d \text{Var}((X_s - Y_s)^2) + \sum_{s=1}^d \sum_{t=1, t \neq s}^d \text{Cov}\left((X_s - Y_s)^2, (X_t - Y_t)^2\right) \right]$

Since the measurement variables from all classes have uniformly bounded fourth moments (see (A1)), we have $\sum_{s=1}^d \text{Var}((X_s - Y_s)^2) = O(d)$. Also, one can show that under assumptions

(A1) and (A2), $\sum_{s=1}^d \sum_{t=1, t \neq s}^d \text{Cov}\left((X_s - Y_s)^2, (X_t - Y_t)^2\right) = o(d^2)$. So, $\text{Var}\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2\right) \rightarrow 0$

and hence $\left|\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2 - E\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2\right)\right| \xrightarrow{P} 0$ as $d \rightarrow \infty$.

Now, $E\left(\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2\right) = E\left(\frac{1}{d}\|(\mathbf{X} - E(\mathbf{X})) - (\mathbf{Y} - E(\mathbf{Y})) + (E(\mathbf{X}) - E(\mathbf{Y}))\|^2\right) = \frac{1}{d}\text{trace}(\boldsymbol{\Sigma}_j) + \frac{1}{d}\text{trace}(\boldsymbol{\Sigma}_i) + \frac{1}{d}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|^2 \rightarrow \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$ as $d \rightarrow \infty$. So, $\frac{1}{d}\|\mathbf{X} - \mathbf{Y}\|^2 \xrightarrow{P} \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$. Note that if \mathbf{X} and \mathbf{Y} follow the same distribution (*i.e.* $j = i$), we have $\nu_{ji}^2 = 0$. \square

Proof of Theorem 1: (a) From Lemma 1, it is clear that for any test case \mathbf{Z} from the j -th class ($j = 1, 2, \dots, J$), $\frac{1}{d}\|\mathbf{Z} - \mathbf{X}_{i\ell}\|^2 \xrightarrow{P} \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$ for $i = 1, 2, \dots, J$ and $\ell = 1, 2, \dots, n_i$. Therefore, for $k < \min\{n_1, n_2, \dots, n_J\}$, the k -nearest neighbor classifier correctly classifies \mathbf{Z} if $2\sigma_j^2 < \sigma_j^2 + \sigma_i^2 + \nu_{ji}^2$ for all $i \neq j$ or equivalently, $\nu_{ji}^2 > \sigma_j^2 - \sigma_i^2$ for all $i \neq j$. Similarly, for correct classification for a test case from the i -th class, we need $\nu_{ij}^2 > \sigma_i^2 - \sigma_j^2$ for all $j \neq i$. Combining these, we get $\nu_{ji}^2 > |\sigma_j^2 - \sigma_i^2|$ for all $j \neq i$.

If $\nu_{ji}^2 < |\sigma_j^2 - \sigma_i^2|$ for any pair (j, i) , we have either $\nu_{ji}^2 < \sigma_j^2 - \sigma_i^2$ or $\nu_{ji}^2 < \sigma_i^2 - \sigma_j^2$. Without loss of generality, let us assume the first one. In that case, for any class- j observation, the distances of its neighbors from the i -th class turn out to be smaller than

those from j -th class with probability tending to 1 as d increases. So, all observations from the j -th class are misclassified with probability tending to 1.

(b) From Lemma 1, it is clear that for any test case \mathbf{Z} from the j -th class, $\frac{1}{d}\rho_j(\mathbf{Z}, \mathbf{X}_{j\ell}) \xrightarrow{P} \sigma_j^2$ for $\ell = 1, 2, \dots, n_j$ whereas for any $i \neq j$, $\frac{1}{d}\rho_i(\mathbf{Z}, \mathbf{X}_{i\ell}) \xrightarrow{P} \sigma_j^2 + \nu_{ji}^2$ for $\ell = 1, 2, \dots, n_i$. So, \mathbf{Z} is correctly classified if $\nu_{ji}^2 > 0$ for all $i \neq j$. Repeating this argument for $j = 1, 2, \dots, J$, we get the result.

(c) Lemma 1 shows that for any test case \mathbf{Z} from the j -th class, $\frac{1}{\sqrt{d}}\rho_j^*(\mathbf{Z}, \mathbf{X}_{j\ell}) \xrightarrow{P} \sigma_j/\sqrt{2}$ for $\ell = 1, 2, \dots, n_j$ whereas for any $i \neq j$, $\frac{1}{\sqrt{d}}\rho_i^*(\mathbf{Z}, \mathbf{X}_{i\ell}) \xrightarrow{P} \sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2} - \sigma_i/\sqrt{2}$ for $\ell = 1, 2, \dots, n_i$. So, \mathbf{Z} is correctly classified if $\sqrt{\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2} > (\sigma_j + \sigma_i)/\sqrt{2}$ or $\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2 > (\sigma_j + \sigma_i)^2/2$ for all $i \neq j$. Note that $\sigma_j^2 + \sigma_i^2 + \nu_{ji}^2 - (\sigma_j + \sigma_i)^2/2 = \nu_{ji}^2 + (\sigma_j - \sigma_i)^2/2$, which is positive under the given condition. Now, the proof follows by the repetition of the same argument for $j = 1, 2, \dots, J$. \square

Proof of Theorem 2: For the sake of simplicity, let us prove it for $J = 2$. For $J > 2$, it can be proved similarly. From Lemma 1, for any training sample observation \mathbf{X}_{1i} ($i = 1, 2, \dots, n_1$) from the first class, as d grows to infinity, we have

$$\left(\frac{1}{\sqrt{d}} \min_{1 \leq \ell (\neq i) \leq n_1} \|\mathbf{X}_{1i} - \mathbf{X}_{1\ell}\|, \frac{1}{\sqrt{d}} \min_{1 \leq \ell \leq n_2} \|\mathbf{X}_{1i} - \mathbf{X}_{2\ell}\| \right) \xrightarrow{P} (\sigma_1\sqrt{2}, \sqrt{\sigma_1^2 + \sigma_2^2 + \nu_{12}^2}) = \mathbf{a}_1, \text{ say.}$$

Similarly, for a training sample observation \mathbf{X}_{2i} ($i = 1, 2, \dots, n_2$) from the second class, as d tends to infinity, we have

$$\left(\frac{1}{\sqrt{d}} \min_{1 \leq \ell \leq n_1} \|\mathbf{X}_{2i} - \mathbf{X}_{1\ell}\|, \frac{1}{\sqrt{d}} \min_{1 \leq \ell (\neq i) \leq n_2} \|\mathbf{X}_{2i} - \mathbf{X}_{2\ell}\| \right) \xrightarrow{P} (\sqrt{\sigma_1^2 + \sigma_2^2 + \nu_{12}^2}, \sigma_2\sqrt{2}) = \mathbf{a}_2, \text{ say.}$$

So, if $\mathbf{a}_1 \neq \mathbf{a}_2$, the feature vectors obtained from two sets of training sample observations converge to two distinct points \mathbf{a}_1 and \mathbf{a}_2 , respectively. Now, for any test case \mathbf{Z} , using Lemma 1, it can be shown that as d grows to infinity, $(\frac{1}{\sqrt{d}} \min_{1 \leq i \leq n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|, \frac{1}{\sqrt{d}} \min_{1 \leq i \leq n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|)$ converges in probability to \mathbf{a}_1 and \mathbf{a}_2 for $\mathbf{Z} \sim F_1$ and $\mathbf{Z} \sim F_2$, respectively.

Therefore, for any $\mathbf{Z} \sim F_1$ (respectively, F_2), in the $d_1 - d_2$ space, while the scaled versions of its distances from the feature vectors from Class-1 (respectively, Class-2) converge to 0, those from the feature vectors from Class-2 (respectively, Class-1) converge to $\|\mathbf{a}_1 - \mathbf{a}_2\|$ as d tends to infinity. So, it is correctly classified with probability tending to 1. Therefore, for perfect classification by the MDist classifier, we need \mathbf{a}_1 and \mathbf{a}_2 to be distinct, *i.e.*, $2\sigma_1^2$, $2\sigma_2^2$ and $\sigma_1^2 + \sigma_2^2 + \nu_{12}^2$ cannot be all equal. Note that these three quantities are equal if and only if $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, which cannot happen under the assumptions of Theorem 2. \square

Lemma 2: Suppose that $\mathbf{X}, \mathbf{X}' \sim F_1$ and $\mathbf{Y}, \mathbf{Y}' \sim F_2$ are four independent d -dimensional random vectors with finite first moments. Then, we have

$$2E\|\mathbf{X} - \mathbf{Y}\|_1 - E\|\mathbf{X} - \mathbf{X}'\|_1 - E\|\mathbf{Y} - \mathbf{Y}'\|_1 \geq 0$$

where the equality holds if and only if F_1 and F_2 have identical one-dimensional marginals.

Proof: First note that

$$2E\|\mathbf{X}_1 - \mathbf{Y}_1\|_1 - E\|\mathbf{X}_1 - \mathbf{X}_2\|_1 - E\|\mathbf{Y}_1 - \mathbf{Y}_2\|_1 = \sum_{q=1}^d \left[2E|X_q - Y_q| - E|X_q - X'_q| - E|Y_q - Y'_q| \right].$$

Now, from Baringhaus and Franz (2004), we get

$$2E|X_q - Y_q| - E|X_q - X'_q| - E|Y_q - Y'_q| = 2 \int_{-\infty}^{\infty} \left(F_1^{(q)}(t) - F_2^{(q)}(t) \right)^2 dt,$$

where $F_1^{(q)}$ and $F_2^{(q)}$ are the distribution functions of X_q and Y_q , respectively. Clearly, it is non-negative, and it takes the value 0 if and only if $F_1^{(q)} = F_2^{(q)}$, i.e., X_q and Y_q have the same distribution. This shows that $2E\|\mathbf{X}_1 - \mathbf{Y}_1\|_1 - E\|\mathbf{X}_1 - \mathbf{X}_2\|_1 - E\|\mathbf{Y}_1 - \mathbf{Y}_2\|_1 \geq 0$, where the equality holds if and only if X_q and Y_q have the same distribution for all $q = 1, 2, \dots, d$. \square

Proof of Theorem 3: As in the proof of Theorem 2, here also, for the sake of simplicity, we prove the result for $J = 2$. For $J > 2$, it can be proved similarly.

Consider two random vectors $\mathbf{X} \sim F_j$ and $\mathbf{Y} \sim F_i$ ($1 \leq j, i \leq 2$). Under (A1°) and (A2°), we have $\left| \frac{1}{d} \|\mathbf{X} - \mathbf{Y}\|_1 - E\left(\frac{1}{d} \|\mathbf{X} - \mathbf{Y}\|_1 \right) \right| \xrightarrow{P} 0$ as $d \rightarrow \infty$, and under (A3°), we have $\lim_{d \rightarrow \infty} E\left(\frac{1}{d} \|\mathbf{X} - \mathbf{Y}\|_1\right) = \tau_{ji}$. Lemma 2 shows that $2\tau_{12} - \tau_{11} - \tau_{22} \geq 0$ and under the assumption $\mathcal{E}_{12} > 0$, the equality is ruled out. So, here we have $2\tau_{12} - \tau_{11} - \tau_{22} > 0$, which implies that τ_{11}, τ_{12} and τ_{22} cannot be equal.

Now note that for any training sample observation \mathbf{X}_{1i} ($i = 1, 2, \dots, n_1$) from the first class, as d grows to infinity,

$$\left(\frac{1}{d} \min_{1 \leq \ell (\neq i) \leq n_1} \|\mathbf{X}_{1i} - \mathbf{X}_{1\ell}\|_1, \frac{1}{d} \min_{1 \leq \ell \leq n_2} \|\mathbf{X}_{1i} - \mathbf{X}_{2\ell}\|_1 \right) \xrightarrow{P} (\tau_{11}, \tau_{12}) = \mathbf{a}_1^\circ, \text{ say.}$$

Similarly, for a training sample observation \mathbf{X}_{2i} ($i = 1, 2, \dots, n_2$) from the second class, as d tends to infinity,

$$\left(\frac{1}{d} \min_{1 \leq \ell \leq n_1} \|\mathbf{X}_{2i} - \mathbf{X}_{1\ell}\|_1, \frac{1}{d} \min_{1 \leq \ell (\neq i) \leq n_2} \|\mathbf{X}_{2i} - \mathbf{X}_{2\ell}\|_1 \right) \xrightarrow{P} (\tau_{12}, \tau_{22}), = \mathbf{a}_2^\circ, \text{ say..}$$

Since τ_{11}, τ_{12} and τ_{22} are not equal, we have $\mathbf{a}_1^\circ \neq \mathbf{a}_2^\circ$. So, the feature vectors obtained from two sets of training sample observations converge to two distinct points \mathbf{a}_1° and \mathbf{a}_2° , respectively. For a test case \mathbf{Z} , as d grows to infinity, $\left(\frac{1}{d} \min_{1 \leq i \leq n_1} \|\mathbf{Z} - \mathbf{X}_{1i}\|_1, \frac{1}{d} \min_{1 \leq i \leq n_2} \|\mathbf{Z} - \mathbf{X}_{2i}\|_1 \right)$ converges in probability to \mathbf{a}_1° and \mathbf{a}_2° for $\mathbf{Z} \sim F_1$ and $\mathbf{Z} \sim F_2$, respectively.

Therefore, for any $\mathbf{Z} \sim F_1$ (respectively, F_2), in the $d_1^* - d_2^*$ space, while the scaled versions of its distances from the feature vectors from Class-1 (respectively, Class-2) converge to 0, those from the feature vectors from Class-2 (respectively, Class-1) converge to $\|\mathbf{a}_1^\circ - \mathbf{a}_2^\circ\|$. So, it is correctly classified with probability tending to 1. \square