Journal of Machine Learning Research 23 (2022) 1-41 Submitted 10/20; Revised 5/21; Published 2/22

On Generalizations of Some Distance Based Classifiers for HDLSS Data

Sarbojit Roy SARBOJIT@IITK.AC.IN
Department of Mathematics and Statistics
IIT Kanpur
Kanpur - 208016, India.

Soham Sarkar SOHAM.SARKAR@EPFL.CH
Institut de Mathématiques
École Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland.

Subhajit Dutta DUTTAS@IITK.AC.IN
Department of Mathematics and Statistics
IIT Kanpur
Kanpur - 208016, India.

Anil K. Ghosh AKGHOSH@ISICAL.AC.IN
Theoretical Statistics and Mathematics Unit
Indian Statistical Institute
Kolkata - 700108, India.

Editor: Charles Elkan

Abstract

In high dimension, low sample size (HDLSS) settings, classifiers based on Euclidean distances like the nearest neighbor classifier and the average distance classifier perform quite poorly if differences between locations of the underlying populations get masked by scale differences. To rectify this problem, several modifications of these classifiers have been proposed in the literature. However, existing methods are confined to location and scale differences only, and they often fail to discriminate among populations differing outside of the first two moments. In this article, we propose some simple transformations of these classifiers resulting in improved performance even when the underlying populations have the same location and scale. We further propose a generalization of these classifiers based on the idea of grouping of variables. High-dimensional behavior of the proposed classifiers is studied theoretically. Numerical experiments with a variety of simulated examples as well as an extensive analysis of benchmark data sets from three different databases exhibit advantages of the proposed methods.

Keywords: Block covariance structure, Convergence in probability, HDLSS asymptotics, Hierarchical clustering, Mean absolute difference of distances, Robustness, Scale-adjusted average distances.

1. Introduction

Classification is a common task in machine learning. Given n data points in $\mathbb{R}^d$ belonging to $J(\geq 2)$ classes, the goal of a classifier is to assign a class label to a new data point. In particular, distance based classifiers have gained popularity because they are quite simple, and easy to implement. Well-known classifiers such as the nearest neighbor classifier, the centroid classifier, and the average distance classifier use only the distance between observations to classify a new test case (see, e.g., Hastie et al., 2009; Chan and Hall, 2009). These classifiers also have nice theoretical properties. Under appropriate conditions, misclassification probabilities of these classifiers converge to the Bayes risk (in other words, *Bayes risk consistency*) as the training sample size increases (see, e.g., Devroye et al., 1996).

In today's world, high-dimensional problems are frequently encountered in scientific areas like microarray gene expression studies, medical image analysis, spectral measurements in chemometrics, etc. A distinct characteristic of some of these problems is the presence of a very large number of features (or, data dimension) with a much smaller sample size. In such high dimension, low sample size (HDLSS) situations, Euclidean distance based classifiers face some natural drawbacks due to *distance concentration* (see, e.g., Aggarwal et al., 2001; Francois et al., 2007). In Hall et al. (2005), the authors studied the effect of distance concentration on some popular classifiers based on Euclidean distances such as the centroid classifier and the nearest neighbor classifier, and derived conditions under which these classifiers yield *perfect classification* in the HDLSS setup. We now give some insight into the idea of distance concentration in HDLSS scenarios.

Consider a random sample $\mathscr{X}_j = \{\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}\}$ of size n_j from the j-th population for $1 \leq j \leq J$. We assume that these $n_j(\geq 2)$ observations are independent and identically distributed (i.i.d.) from a distribution function $\mathbf{F}_j$ on $\mathbb{R}^d$. Define $\mathscr{X} = \cup_{j=1}^{J} \mathscr{X}_j$ to be the full training sample of size $n = \sum_{j=1}^{J} n_j$. For simplicity of analysis, we take $J = 2$. Let $\boldsymbol{\mu}_{jd}$ and Σ_{jd} denote the d-dimensional location vector and the $d \times d$ scale matrix, respectively, corresponding to $\mathbf{F}_j$ for $j = 1, 2$. Also, assume that the following limits exist:

$$\nu_{12}^2 := \lim_{d\to\infty} \left\{ d^{-1} \|\boldsymbol{\mu}_{1d} - \boldsymbol{\mu}_{2d}\|^2 \right\} \text{ and } \sigma_j^2 = \lim_{d\to\infty} \left\{ d^{-1} \mathrm{tr}(\Sigma_{jd}) \right\} \text{ for } j = 1, 2.$$

Here, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$ and $\mathrm{tr}(A)$ is the sum of the diagonal elements of a $d \times d$ matrix A. The constants ν_{12}^2 and σ_1^2, σ_2^2 are measures of the location difference and scales, respectively. In Hall et al. (2005), the authors showed that in the HDLSS asymptotic regime (when n is fixed and d goes to infinity), if $\nu_{12}^2 < |\sigma_1^2 - \sigma_2^2|$, the nearest neighbor (NN) classifier *assigns all observations to the population having a smaller dispersion.* Later, Chan and Hall (2009) showed that the average distance (AVG) classifier is also *useless* in such a scenario. In other words, Euclidean distance based classifiers may not yield satisfactory performance for high-dimensional data if the location difference is masked by the scale difference. To address this specific problem, some modifications of these classifiers have been proposed in the literature. Chan and Hall (2009) identified $|\sigma_1^2 - \sigma_2^2|$ as a nuisance parameter, and proposed a scale adjustment to the discriminant of the average distance classifier. A non-linear transformation of the covariate space followed by NN classification was proposed by Dutta and Ghosh (2016), while Pal et al. (2016) developed a NN classifier based on a new dissimilarity index. However, all these modified classifiers are known to perform well in the HDLSS setup under conditions like '$\nu_{12}^2 > 0$' or 'either $\nu_{12}^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$'. To summarize, all the existing classifiers are particularly useful in high-dimensional spaces when the underlying distributions differ either in their locations and/or scales. Our interest is to analyze the performance of these classifiers under more general scenarios (in particular, when $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$). We demonstrate this by considering some classification problems involving two populations.

Example 1 *We consider two populations where the d component variables are i.i.d. For the first population, the component distribution is $N(0, 5/3)$, while it is t_5 for the second population. Here, $N(\mu, \sigma^2)$ denotes the univariate Gaussian distribution with mean μ and variance σ^2, and t_ν denotes the standard Student's t distribution with ν degrees of freedom.*

Example 2 *The two populations under consideration have the d-dimensional Gaussian distributions $N_d(\mathbf{0}_d, \Sigma_{1d})$ and $N_d(\mathbf{0}_d, \Sigma_{2d})$, where $\mathbf{0}_d$ is the d-dimensional vector of zeros, and Σ_{1d} and Σ_{2d} are block diagonal dispersion matrices having the following form:*

$$\Sigma_{jd} = \begin{bmatrix} \mathbf{H}_j & 0 & \cdots & 0 \\ 0 & \mathbf{H}_j & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{H}_j \end{bmatrix} \textit{ with } \mathbf{H}_j = \begin{bmatrix} 1 & \rho_j \cdots \rho_j \\ \rho_j & 1 \cdots \rho_j \\ \vdots & \vdots \ddots \vdots \\ \rho_j & \rho_j \cdots 1 \end{bmatrix} \textit{ for } j = 1, 2.$$

In this example, we keep the size of the blocks fixed at ten (i.e., $\mathbf{H}_j$ is a 10×10 matrix for $j = 1, 2$) and choose $\rho_1 = 0.3$ and $\rho_2 = 0.7$.

Example 3 *We consider d-dimensional Gaussian distributions $N_d(\mathbf{0}_d, \Sigma_{1d})$ and $N_d(\mathbf{0}_d, \Sigma_{2d})$, where Σ_{1d} and Σ_{2d} have an auto-regressive covariance structure (i.e., $\Sigma_d = ((\rho^{|i-j|}))_{1 \leq i,j \leq d}$ and $0 < \rho < 1$) with parameters 0.3 and 0.7, respectively.*

For each example, we generated 50 observations from each class to form the training sample. Misclassification rates of different classifiers are computed based on a test set consisting of 500 (250 from each class) observations. This process was repeated 100 times, and the average misclassification rates (along with the standard errors) of different classifiers for varying values of d are shown in Figure 1. The Bayes risk was calculated for each example by computing the average Bayes risk over several random replicates of the data. It is clear from Figure 1 that none of the existing classifiers performed satisfactorily in these three examples. Observe that in all three examples, we have $\nu_{12}^2 = 0$ (the mean vectors $\boldsymbol{\mu}_{1d}$ and $\boldsymbol{\mu}_{2d}$ are equal to $\mathbf{0}_d$) and $\sigma_1^2 = \sigma_2^2$ (both Σ_{1d} and Σ_{2d} have the same trace). This was the main reason behind the poor performance of all the existing classifiers.

In this article, we propose a modification to the Euclidean distance, and use it on two different distance based classifiers, namely, the scale-adjusted average distance classifier (henceforth referred to as SAVG) by Chan and Hall (2009) and the NN classifier based on mean absolute differences of distances (henceforth referred to as NN-MADD) by Pal et al. (2016). We show that these two classifiers, when used with the modified distance, can discriminate between populations even when there are no differences between their locations and scales. To capture discriminatory information, these modified distance based classifiers rely on the non-parametric concept of *energy* (see Székely and Rizzo, 2017). In particular, if the one-dimensional marginals of the underlying populations are different, the proposed

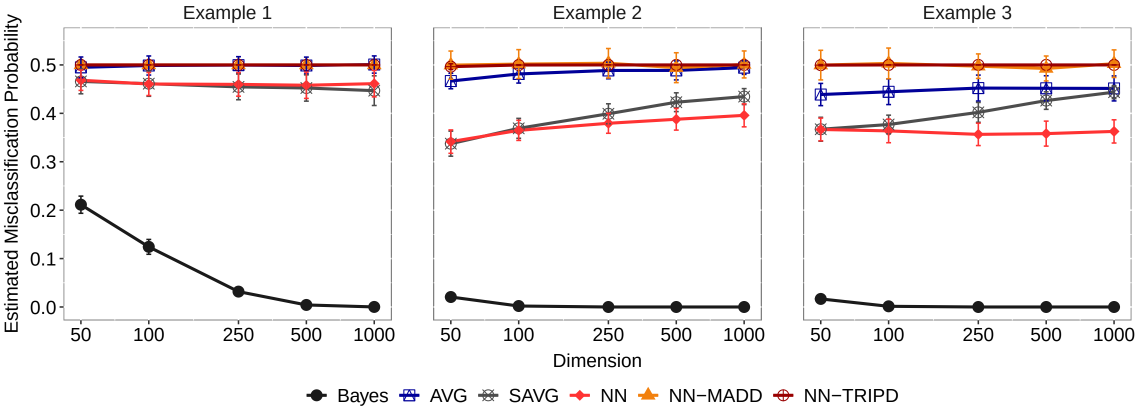

Figure 1: Average misclassification rates (along with the standard errors) based on 100 repetitions of various classifiers are plotted for increasing values of d (in logarithmic scale). The classifiers AVG/SAVG, NN-MADD and NN-TRIPD were proposed by Chan and Hall (2009), Pal et al. (2016) and Dutta and Ghosh (2016), respectively.

classifiers are shown to yield *perfect classification* in the HDLSS asymptotic regime. For HDLSS asymptotics, we fix the sample size n and allow the data dimension d to grow to infinity, which is different from standard asymptotics (with d fixed and n going to infinity).

The article is organized as follows. We define the modified classifiers and study their asymptotic properties in Section 2. In Section 3, we propose further generalization of these classifiers for the case when the populations have same univariate marginals, but differ in their joint distributional structures (see Examples 2 and 3) and derive their asymptotic properties under the HDLSS setup. For implementation of the second generalization, we need to group the component variables into disjoint clusters. In Section 4, we propose some data driven methods for this 'variable clustering'. Numerical performance of the proposed classifiers on several simulated and real data sets are demonstrated in Sections 5 and 6, respectively. The article ends with a discussion in Section 7. All proofs and other mathematical details are provided in Appendix A, and some additional material is presented as a Supplementary. A list of notations used in this paper is given in Appendix B.

2. Classifiers Based on Generalized Distances

Limitations of the classifiers discussed in the previous section stems from the fact that the behavior of the Euclidean distance in the HDLSS asymptotic regime is completely governed by the constants ν_{12}^2, σ_1^2 and σ_2^2 (see Hall et al., 2005). As a consequence, Euclidean distance based classifiers cannot distinguish between populations that do not have differences in their first two moments. To circumvent this problem, we define a class of dissimilarity measures. For vectors $\mathbf{u} = (u_1, \ldots, u_d)^\top$ and $\mathbf{v} = (v_1, \ldots, v_d)^\top$, we define the dissmimilarity function $h_d^{\phi,\gamma} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ between $\mathbf{u}$ and $\mathbf{v}$ as follows:

$$h_d^{\phi,\gamma}(\mathbf{u}, \mathbf{v}) \equiv h_d(\mathbf{u}, \mathbf{v}) = \phi\biggl(\frac{1}{d}\sum_{i=1}^{d} \gamma\bigl(|u_i - v_i|^2\bigr)\biggr), \tag{2.1}$$

where $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ and $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ are continuous, monotonically increasing with $\gamma(0) = \phi(0) = 0$. The class of functions (2.1) was proposed and used in the context of two-sample testing in Sarkar and Ghosh (2018). It is interesting to note that if $\gamma(t) = t^{p/2}$ and $\phi(t) = t^{1/p}$ with $p > 0$, then $h_d(\mathbf{u}, \mathbf{v})$ is the ℓ_p distance (up to a constant involving d) between $\mathbf{u}$ and $\mathbf{v}$. This in particular includes the Euclidean distance (for $p = 2$) as a special case. In general, $h_d(\mathbf{u}, \mathbf{v})$ need not be a distance function, but rather a measure of dissimilarity between $\mathbf{u}$ and $\mathbf{v}$. Our main objective is to use $h_d(\mathbf{u}, \mathbf{v})$ instead of the *scaled* Euclidean distance (i.e., $d^{-1}\|\mathbf{u} - \mathbf{v}\|^2$ or $d^{-1/2}\|\mathbf{u} - \mathbf{v}\|$) in the SAVG and NN-MADD classifiers, and study their performance, both theoretically as well as numerically.

2.1 Generalization of SAVG Classifier

For a J-class problem and a new observation $\mathbf{Z}$, the average distance (AVG) classifier is defined as

$$\delta_{\text{AVG}}(\mathbf{Z}) = \underset{1 \le j \le J}{\arg\min} \left\{ \frac{1}{n_j} \sum_{\mathbf{X} \in \mathscr{X}_j} d^{-1}\|\mathbf{X} - \mathbf{Z}\|^2 \right\}. \tag{2.2}$$

If $\nu^2_{jj'} > |\sigma^2_j - \sigma^2_{j'}|$ for all $1 \le j \ne j' \le J$, then this classifier yields perfect classification in the HDLSS setup (i.e., the misclassification probability of the classifier goes to zero as $d \to \infty$, see Chan and Hall, 2009). But, if this condition is violated, then this classifier may behave erratically by assigning all observations to the class having the smallest variance. To relax the condition stated above, the authors identified $|\sigma^2_j - \sigma^2_{j'}|$ as a nuisance parameter, and proposed a scale adjustment to the average of distances as follows:

$$\xi^{(0)}_{jd}(\mathbf{Z}) = \frac{1}{n_j} \sum_{\mathbf{X} \in \mathscr{X}_j} d^{-1}\|\mathbf{X} - \mathbf{Z}\|^2 - D^{(0)}_d(\mathscr{X}_j|\mathscr{X}_j)/2, \tag{2.3}$$

where $D^{(0)}_d(\mathscr{X}_j|\mathscr{X}_j) = \{n_j(n_j - 1)\}^{-1} \sum_{\mathbf{X}, \mathbf{X}' \in \mathscr{X}_j} d^{-1}\|\mathbf{X} - \mathbf{X}'\|^2$ for all $1 \le j \le J$. The scale-adjusted average distance (SAVG) classifier is defined as

$$\delta_{\text{SAVG}}(\mathbf{Z}) = \underset{1 \le j \le J}{\arg\min}\, \xi^{(0)}_{jd}(\mathbf{Z}).$$

If $\nu^2_{jj'} > 0$ for all $1 \le j \ne j' \le J$, then the misclassification probability of the SAVG classifier goes to zero as $d \to \infty$ (see Chan and Hall, 2009, Theorem 1). The optimality condition for the SAVG classifier is clearly weaker than the one related to the AVG classifier. In other words, if the competing populations have difference only in their location parameters (irrespective of their differences in scales), the SAVG classifier perfectly classifies a new data point in high dimensions. However, we have observed deteriorating performance of the SAVG classifier in Figure 1 when this condition is violated (recall that $\nu^2_{12} = 0$ in Examples 1, 2 and 3).

We modify the SAVG classifier by simply replacing the Euclidean distance $d^{-1}\|\mathbf{u} - \mathbf{v}\|^2$ with the new dissimilarity index $h_d(\mathbf{u}, \mathbf{v})$, as stated below:

$$\xi^{\phi,\gamma}_{jd}(\mathbf{Z}) \equiv \xi_{jd}(\mathbf{Z}) = \frac{1}{n_j} \sum_{\mathbf{X} \in \mathscr{X}_j} h_d(\mathbf{Z}, \mathbf{X}) - D_d(\mathscr{X}_j|\mathscr{X}_j)/2. \tag{2.4}$$

Here, $D_d(\mathscr{X}_j|\mathscr{X}_j) = \{n_j(n_j-1)\}^{-1}\sum_{\mathbf{X},\mathbf{X}'\in\mathscr{X}_j} h_d(\mathbf{X},\mathbf{X}')$ with $n_j \geq 2$ for $1 \leq j \leq J$. The generalized scale-adjusted average distance (gSAVG) classifier based on ξ_{jd} is given by

$$\delta_{\text{gSAVG}}(\mathbf{Z}) = \underset{1\leq j\leq J}{\arg\min}\, \xi_{jd}(\mathbf{Z}). \tag{2.5}$$

Observe that ξ_{jd} reduces to the earlier transformation $\xi_{jd}^{(0)}$ if we consider $\gamma(t) = t$ and $\phi(t) = t$ in equation (2.1). So, the gSAVG classifier is a generalization of the SAVG classifier.

2.2 Generalization of NN-MADD Classifier

For a test point $\mathbf{Z} \in \mathbb{R}^d$, the usual nearest neighbor (NN) classifier is defined as follows:

$$\delta_{\text{NN}}(\mathbf{Z}) = \underset{1\leq j\leq J}{\arg\min}\, \tau_{jd}(\mathbf{Z}), \tag{2.6}$$

where $\tau_{jd}(\mathbf{Z}) = \min_{\mathbf{X}\in\mathscr{X}_j} \|\mathbf{Z}-\mathbf{X}\|$ for $1 \leq j \leq J$. In high dimensions, the NN classifier perfectly classifies a new observation when $\nu_{jj'}^2 > |\sigma_j^2 - \sigma_{j'}^2|$ for all $1 \leq j \neq j' \leq J$ (see Hall et al., 2005). But, when this condition is violated, this classifier may behave erratically (see, e.g., Pal et al., 2016). To avoid this problem, Pal et al. (2016) proposed an approach by modifying the distance function and defined the dissimilarity between $\mathbf{Z}$ and a training observation $\mathbf{X} \in \mathscr{X}$ as follows:

$$\psi_d^{(0)}(\mathbf{Z},\mathbf{X}) = \frac{1}{n-1}\sum_{\mathbf{X}'\in\mathscr{X}\setminus\mathbf{X}} \Big| d^{-1/2}\|\mathbf{Z}-\mathbf{X}'\| - d^{-1/2}\|\mathbf{X}-\mathbf{X}'\| \Big|. \tag{2.7}$$

The dissimilarity $\psi_d^{(0)}$ is called the mean absolute difference of distances (MADD). The NN classifier based on MADD is defined as

$$\delta_{\text{NN}-\text{MADD}}(\mathbf{Z}) = \underset{1\leq j\leq J}{\arg\min}\, \tau_{jd}^{(0)}(\mathbf{Z}), \tag{2.8}$$

where $\tau_{jd}^{(0)}(\mathbf{Z}) = \min_{\mathbf{X}\in\mathscr{X}_j} \psi_d^{(0)}(\mathbf{Z},\mathbf{X})$ for $1 \leq j \leq J$. The NN-MADD classifier perfectly classifies a new observation in the HDLSS setup when $\nu_{jj'}^2 > 0$ or $\sigma_j^2 \neq \sigma_{j'}^2$ for all $1 \leq j \neq j' \leq J$. This condition is clearly weaker than the one for the usual NN classifier stated above. However, this classifier too performed quite poorly in Examples 1, 2 and 3, where the condition was violated.

Here again, the problem lies in the use of Euclidean distance in the construction of $\psi_d^{(0)}$. To resolve this issue, we use the new distance function h_d defined in (2.1) to modify the transformation $\psi_d^{(0)}$ given in (2.7) as follows:

$$\psi_d^{\phi,\gamma}(\mathbf{Z},\mathbf{X}) \equiv \psi_d(\mathbf{Z},\mathbf{X}) = \frac{1}{n-1}\sum_{\mathbf{X}'\in\mathscr{X}\setminus\mathbf{X}} \big| h_d(\mathbf{Z},\mathbf{X}') - h_d(\mathbf{X},\mathbf{X}') \big|. \tag{2.9}$$

The dissimilarity index ψ_d is referred to as mean absolute difference of generalized distances (or, generalized MADD and hence, abbreviated as gMADD). Using gMADD, we define

$\tau_{jd}(\mathbf{Z}) = \min_{\mathbf{X}\in\mathscr{X}_j} \psi_d(\mathbf{Z},\mathbf{X})$ for $1 \le j \le J$. The associated nearest neighbor classifier is defined as

$$\delta_{\mathrm{NN-gMADD}}(\mathbf{Z}) = \underset{1\le j\le J}{\arg\min}\, \tau_{jd}(\mathbf{Z}). \tag{2.10}$$

If we consider $\gamma(t) = t$ and $\phi(t) = \sqrt{t}$ in (2.1), then ψ_d reduces to $\psi_d^{(0)}$ defined in (2.7). Consequently, the NN-gMADD classifier reduces to the NN-MADD classifier.

Recall that in Examples 1, 2 and 3 we have $\nu_{12}^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. So, both the classifiers SAVG and NN-MADD (based on Euclidean distances) performed quite poorly (see Figure 1). However, Figure 2 clearly shows the superiority of the proposed gSAVG and NN-gMADD classifiers in Example 1 with $\gamma(t) = 1 - e^{-t}$ and $\phi(t) = t$. In high dimensions, they have misclassification rates close to the Bayes risk. The misclassification rates of different NN classifiers are reported by considering a single neighbor (i.e., for $k = 1$) only. We observed a similar phenomenon for other values of k as well. In Figure 2, we further observe that both the gSAVG and NN-gMADD classifiers misclassify nearly 50% and 45% (for higher values of d) of the test samples in Examples 2 and 3, respectively. Interestingly, the transformation h_d works favourably for Example 1, while it is quite intriguing to note that it fails to yield good performance in Examples 2 and 3 for high d. In the next subsection, we study the reason behind this behavior of the proposed classifiers in high dimensions. We begin by studying the theoretical behavior of the transformation h_d in the HDLSS asymptotic regime.

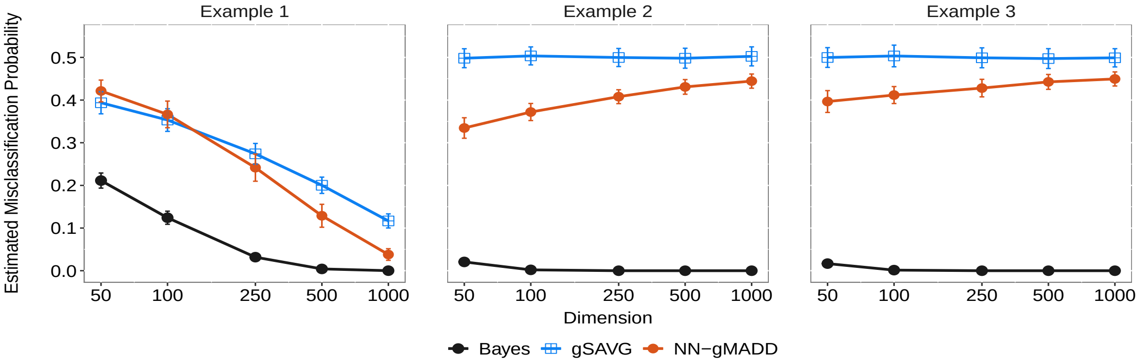

Figure 2: Average misclassification rates (along with the standard errors) based on 100 repetitions of the gSAVG and NN-gMADD classifiers are plotted for increasing values of d (in logarithmic scale).

2.3 Behavior of Generalized Classifiers in HDLSS Asymptotic Regime

Suppose that $\mathbf{U} = (U_1, \ldots, U_d)^\top \sim \mathbf{F}_j$ and $\mathbf{V} = (V_1, \ldots, V_d)^\top \sim \mathbf{F}_{j'}$ are two independent d-dimensional random vectors. We denote the marginal distribution of the i-th component corresponding to the j-th population by $F_{j,i}$ for $1 \le i \le d$ and $1 \le j \le J$. To study the asymptotic behavior of $h_d^{\phi,\gamma}$, we make the following assumptions:

$(A1)$ There exists a constant c_1 such that $\mathrm{E}\big(\gamma^2(|U_i - V_i|^2)\big) \le c_1 < \infty\ \forall\ 1 \le i \le d$.

$(A2)$ $\sum\sum_{1\le i<i'\le d} \mathrm{Corr}\big(\gamma(|U_i - V_i|^2), \gamma(|U_{i'} - V_{i'}|^2)\big) = o(d^2)$.

It is evident that $(A1)$ is satisfied if γ is bounded. Assumption $(A2)$ holds if the component variables of the underlying populations are independent. However, it continues to hold even when the components are dependent, with some additional conditions on their dependence structure. For instance, in the case of sequence data, $(A2)$ holds when the sequence has the ρ-mixing property (see, e.g., Hall et al., 2005; Bradley, 2005). Conditions similar to $(A2)$ have been considered previously for studying the high-dimensional behavior of different statistical methods (see the review paper by Aoshima et al., 2018). Under assumptions $(A1)$ and $(A2)$, the high-dimensional behavior of $h_d^{\phi,\gamma}$ is given by the following lemma.

Lemma 2.1 *Suppose that $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$ are two independent random vectors satisfying assumptions $(A1)$ and $(A2)$ with $1 \leq j, j' \leq J$, and ϕ is uniformly continuous. Then*

$$\left|h_d(\mathbf{U}, \mathbf{V}) - \tilde{h}_d(j, j')\right| \xrightarrow{P} 0 \text{ as } d \to \infty,$$

where $\tilde{h}_d(j, j') \equiv \tilde{h}_d^{\phi,\gamma}(j, j')$ is defined as $\tilde{h}_d(j, j') = \phi[d^{-1} \sum_{i=1}^{d} \mathrm{E}\{\gamma(|U_i - V_i|^2)\}]$.

For $1 \leq j, j' \leq J$, define the following quantities:

$$\tilde{\xi}_d^{\phi,\gamma}(j, j') \equiv \tilde{\xi}_d(j, j') = \tilde{h}_d(j, j') - \frac{1}{2}\big[\tilde{h}_d(j', j') + \tilde{h}_d(j, j)\big], \text{ and}$$

$$\tilde{\tau}_d^{\phi,\gamma}(j, j') \equiv \tilde{\tau}_d(j, j') = \sum_{1 \leq l \neq j' \leq J} \left[\frac{n_l}{n-1} |\tilde{h}_d(j', l) - \tilde{h}_d(j, l)|\right] + \frac{n_{j'} - 1}{n-1} |\tilde{h}_d(j', j') - \tilde{h}_d(j, j')|.$$

As an immediate consequence of Lemma 2.1, we get the following result involving $\xi_{jd}(\mathbf{Z})$ (defined in (2.4)) and $\tau_{jd}(\mathbf{Z})$ (defined just above (2.10)).

Corollary 2.2 *If a test observation $\mathbf{Z} \sim \mathbf{F}_j$, then for any $1 \leq j' \leq J$ we have*

(a) $\left|\big\{\xi_{j'd}(\mathbf{Z}) - \xi_{jd}(\mathbf{Z})\big\} - \tilde{\xi}_d(j, j')\right| \xrightarrow{P} 0$ *as* $d \to \infty$,

(b) $\left|\{\tau_{j'd}(\mathbf{Z}) - \tau_{jd}(\mathbf{Z})\} - \tilde{\tau}_d(j, j')\right| \xrightarrow{P} 0$ *as* $d \to \infty$.

From the definition, it is clear that $\tilde{\xi}_d$ is symmetric (i.e., $\tilde{\xi}_d(j, j') = \tilde{\xi}_d(j', j)$) and $\tilde{\xi}_d(j, j) = 0$ for $1 \leq j, j' \leq J$. Recall that δ_{gSAVG} classifies $\mathbf{Z} \sim \mathbf{F}_j$ correctly if $\xi_{j'd}(\mathbf{Z}) - \xi_{jd}(\mathbf{Z}) > 0$ for all $j' \neq j$. So, for good performance of gSAVG in high dimensions, it is expected that we have $\tilde{\xi}_d(j, j') > 0$ for large values of d. On the other hand, the constant $\tilde{\tau}_d(j, j')$ is non-negative and $\tilde{\tau}_d(j, j) = 0$ for all $1 \leq j, j' \leq J$ by definition. Again, it is desirable to have $\tilde{\tau}_d(j, j') > 0$ for large values of d, to ensure good performance of the NN-gMADD classifier. Both these requirements are met by choosing the functions ϕ and γ appropriately, as stated in the following lemma.

Lemma 2.3 *Let γ have non-constant, completely monotone derivative on $\mathbb{R}^+$. Then, the following results hold.*
(a) If ϕ is concave, then $\tilde{\xi}_d(j, j') \geq 0$, and $\tilde{\xi}_d(j, j') = 0$ if and only if $F_{j,i} = F_{j',i}$ for all $1 \leq i \leq d$.
(b) If ϕ is one-to-one, then $\tilde{\tau}_d(j, j') = 0$ if and only if $F_{j,i} = F_{j',i}$ for all $1 \leq i \leq d$.

Functions with non-constant, completely monotone derivatives have been considered earlier in the literature (see, e.g., Feller, 1971; Baringhaus and Franz, 2010). Lemma 2.3 shows that for appropriate choices of ϕ and γ, the quantity $\tilde{\xi}_d(j,j')$ can be viewed as a measure of separation between the two population distribution functions $\mathbf{F}_j$ and $\mathbf{F}_{j'}$ for $1 \leq j \neq j' \leq J$. In fact, this quantity attains the value zero only when the two populations have identical one-dimensional marginals, and it is related to the idea of *energy* (see Székely and Rizzo, 2017). So, it is reasonable to assume the following:

$(A3)$ For every $1 \leq j \neq j' \leq J$, $\liminf\limits_{d\to\infty} \tilde{\xi}_d(j,j') > 0$.

This assumption ensures that separation among the populations is asymptotically non-negligible. A similar condition for $\tilde{\tau}_d(j,j')$ follows from assumption $(A3)$ (see Lemma 1 in Appendix A). The following theorem states the high-dimensional behavior of the proposed classifiers under these assumptions.

Theorem 2.4 *Define $n_0 = \min\{n_1,\ldots,n_J\}$. If assumptions (A1)–(A3) are satisfied, then*
(a) for any $n_0 \geq 2$, the misclassification probability of the gSAVG classifier converges to zero as $d \to \infty$, and
(b) for any $k \leq n_0$, the misclassification probability of the k-NN classifier based on gMADD converges to zero as $d \to \infty$.

When the underlying distributions have different marginal distributions, Theorem 2.4 suggests that classifiers based on the transformation $h_d^{\phi,\gamma}$ should have excellent performance if ϕ and γ are chosen appropriately. The choice $\phi(t) = t$ satisfies the conditions of Lemmas 2.1 and 2.3. There are several choices of γ that satisfy the conditions stated in Lemma 2.3 (see Baringhaus and Franz, 2010, p.1338). In particular, $\gamma(t) = 1 - e^{-t}$ satisfies these conditions.

Let us now recall Figure 2. In Example 1, the one-dimensional marginals of $\mathbf{F}_1$ are all $N(0, 5/3)$, while for $\mathbf{F}_2$ the marginals are t_5. So, there is difference in the one-dimensional marginal distributions and assumptions $(A1) - (A3)$ are satisfied in this example. On the other hand, the marginal distributions of both classes are same (namely, $N(0,1)$) in Examples 2 and 3. As a result, assumption $(A3)$ is violated and Theorem 2.4 fails to hold in these two examples.

3. Further Generalization Using Groups of Variables

In Figure 2, we have observed that the proposed classifiers fail to discriminate among populations for which the one-dimensional marginals are identical (recall Examples 2 and 3). However, in Example 2 we have information in 'groups of variables' and the groups are quite prominent. If we can capture this information in the joint structure of the sub-vectors (instead of extracting information only from the d univariate components) and modify our classifiers accordingly, it is expected that the classifiers will perform better. In this section, we use this idea to further generalize the transformations $\xi_d^{\phi,\gamma}$ and $\tau_d^{\phi,\gamma}$ so that populations can be discriminated even when the one-dimensional marginals are same.

To build the next step of generalization, we assume that the component variables of a high-dimensional vector have an implicit property of forming groups of variables. By groups of variables, we simply mean a non-overlapping collection of variables. We will address the problem of finding these groups in practice later in Section 4. Meanwhile, let us assume that the groups are known, i.e., the components of a d-dimensional vector $\mathbf{u}$ are partitioned into b known groups. Let $\mathcal{C} = \{C_1, \ldots, C_b\}$ represent the collection of these groups, where $C_i = \{l_{d_{i-1}+1}, \ldots, l_{d_i}\}$ with $d_0 = 0$ and $1 \leq i \leq b$. Now, consider the sub-vector $\mathbf{u}_i = (u_{l_{d_{i-1}+1}}, \ldots, u_{l_{d_i}})^\top$ of dimension d_i for $1 \leq i \leq b$. We propose a modification of $h_d^{\phi,\gamma}$ so that the discriminants can extract information from the distributions of these sub-vectors (i.e., groups of component variables).

For two vectors $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_b^\top)^\top$ and $\mathbf{v} = (\mathbf{v}_1^\top, \ldots, \mathbf{v}_b^\top)^\top$, we define a generalized dissimilarity measure as follows:

$$h_b^{\phi,\gamma}(\mathbf{u}, \mathbf{v}) \equiv h_b(\mathbf{u}, \mathbf{v}) = \phi\left[\frac{1}{b}\sum_{i=1}^{b} \gamma\left(d_i^{-1}\|\mathbf{u}_i - \mathbf{v}_i\|^2\right)\right]. \tag{3.1}$$

We would like to point out the notational similarity between equations (3.1) and (2.1). Throughout the article, we use the convention that with suffix d, we denote the generalized distance based on component variables as defined in (2.1), while with suffix b, we denote the generalized distance based on groups of variables as defined in (3.1).

We first modify the gSAVG classifier defined in (2.5) as follows. Using the transformation $h_b^{\phi,\gamma}$, we define

$$\xi_{jb}^{\phi,\gamma}(\mathbf{Z}) \equiv \xi_{jb}(\mathbf{Z}) = \frac{1}{n_j}\sum_{\mathbf{X}\in\mathscr{X}_j} h_b(\mathbf{Z}, \mathbf{X}) - D_b(\mathscr{X}_j|\mathscr{X}_j)/2, \tag{3.2}$$

where $D_b(\mathscr{X}_j|\mathscr{X}_j) = \{n_j(n_j-1)\}^{-1}\sum_{\mathbf{X},\mathbf{X}'\in\mathscr{X}_j} h_b(\mathbf{X}, \mathbf{X}')$ for $1 \leq j \leq J$. Now, the block-generalized SAVG (bgSAVG) classifier is defined as

$$\delta_{\text{bgSAVG}}(\mathbf{Z}) = \underset{1\leq j\leq J}{\arg\min}\, \xi_{jb}(\mathbf{Z}). \tag{3.3}$$

Similarly, we modify the NN-gMADD classifier defined in (2.10) as follows. Define

$$\psi_b^{\phi,\gamma}(\mathbf{Z}, \mathbf{X}) \equiv \psi_b(\mathbf{Z}, \mathbf{X}) = \frac{1}{n-1}\sum_{\mathbf{X}'\in\mathscr{X}\setminus\mathbf{X}} \left|h_b(\mathbf{Z}, \mathbf{X}') - h_b(\mathbf{X}, \mathbf{X}')\right|, \tag{3.4}$$

and $\tau_{jb}(\mathbf{Z}) = \min_{\mathbf{X}\in\mathscr{X}_j} \psi_b(\mathbf{Z}, \mathbf{X})$ for $1 \leq j \leq J$. The associated nearest neighbor classifier is now defined as:

$$\delta_{\text{NN-bgMADD}}(\mathbf{Z}) = \underset{1\leq j\leq J}{\arg\min}\, \tau_{jb}(\mathbf{Z}). \tag{3.5}$$

We refer to $\delta_{\text{NN-bgMADD}}$ as the NN classifier based on block-generalized MADD (or, the NN-bgMADD classifier).

Let us now investigate the performance of the proposed classifiers in Examples 2 and 3. The choice of groups is quite clear in Example 2 (we have $d_i = 10$ for all $1 \leq i \leq b$ with $C_1 = \{1, \ldots, 10\}$; $C_2 = \{11, \ldots, 20\}$; and so on), but it is not so straightforward in Example 3. In both examples, we formed equal-sized groups using consecutive variables with varying choices of the group sizes, and the corresponding results are shown in Figure 3.

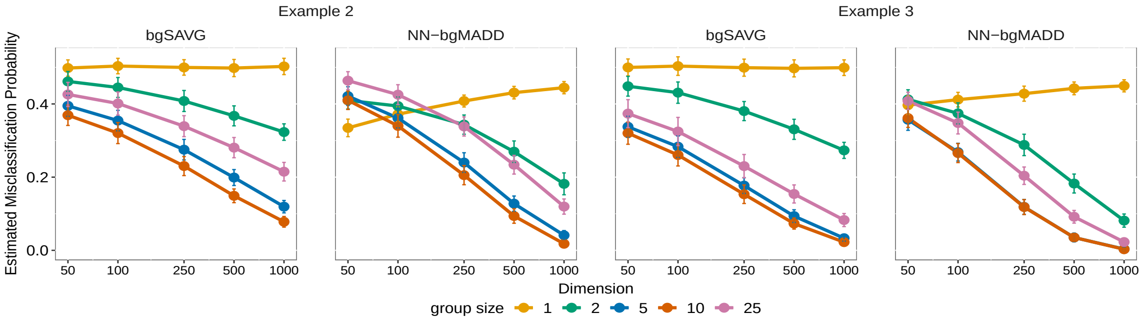

Figure 3: Average misclassification rates (along with the standard errors) based on 100 repetitions of the bgSAVG and NN-bgMADD classifiers are plotted with varying group sizes for increasing values of d (in logarithmic scale).

Figure 3 clearly shows the superiority of the modified (both bgSAVG and NN-bgMADD with $\gamma(t) = 1 - e^{-t}$ and $\phi(t) = t$) classifiers when compared with the gSAVG and NN-gMADD (i.e., $d_i = 1$ for all $1 \leq i \leq d$) classifiers. In high dimensions, the block-generalized classifiers have misclassification rates quite close to zero (even for low values of d_i like 5). On the other hand, the performance deteriorates when the value of d_i is increased to 25. Clearly, this reflects that the choice of group size is quite crucial for the proposed classifiers to perform well in practice. We provide details on the practical implementation of variable clustering for the block-generalized classifiers in Section 4. But first, we study the theoretical behavior of h_b and the two associated classifiers, viz., bgSAVG and NN-bgMADD in the HDLSS asymptotic regime.

3.1 Behavior of Block-Generalized Classifiers in HDLSS Asymptotic Regime

Recall that the HDLSS asymptotic behavior of the generalized distance h_d (and associated classifiers) depend on the one-dimensional marginal distributions $F_{j,i}$ for $1 \leq i \leq d$ and $1 \leq j \leq J$. Similarly, the HDLSS asymptotic behavior of h_b (and related classifiers) will be governed by the joint distributions of groups of variables. To this extent, let us assume that we have a common cluster structure $\mathcal{C}$ along all the J classes, and $\mathcal{C}$ is known. For a random vector $\mathbf{U} = (\mathbf{U}_1^\top, \ldots, \mathbf{U}_b^\top)^\top \sim \mathbf{F}_j$ partitioned according to $\mathcal{C}$, we denote the distribution function of $\mathbf{U}_i$ by $\mathbf{F}_{j,i}$ for $1 \leq i \leq b$ and $1 \leq j \leq J$. To study the HDLSS asymptotic behavior of the newly proposed classifiers (viz., bgSAVG and NN-bgMADD), we restrict ourselves to the setting where the sizes of clusters d_i remain bounded for $1 \leq i \leq b$. This assumption is formally stated below.

$(A4)$ There exists a fixed positive integer d_0 such that $d_i \leq d_0$ for all $1 \leq i \leq b$.

It is clear from assumption $(A4)$ that $b \leq d = \sum_{i=1}^{b} d_i \leq bd_0$. Hence, we can write '$b \to \infty$' and '$d \to \infty$' interchangeably. Now, for $\mathbf{U} = (\mathbf{U}_1^\top, \ldots, \mathbf{U}_b^\top)^\top \sim \mathbf{F}_j$ and $\mathbf{V} = (\mathbf{V}_1^\top, \ldots, \mathbf{V}_b^\top)^\top \sim \mathbf{F}_{j'}$ with $1 \leq j, j' \leq J$, consider the following assumptions:

$(A5)$ There exists a constant c_2 such that $\mathrm{E}[\gamma^2(d_i{}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2)] \leq c_2$ for all $1 \leq i \leq b$.

$(A6)$ $\sum\sum_{1 \leq i < i' \leq b} \mathrm{Corr}\big[\gamma\big(d_i{}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2\big), \gamma\big(d_{i'}{}^{-1}\|\mathbf{U}_{i'} - \mathbf{V}_{i'}\|^2\big)\big] = o(b^2)$.

Assumptions $(A5)$ and $(A6)$ are generalizations of assumptions $(A1)$ and $(A2)$, respectively. As we observed earlier, choosing γ to be bounded is sufficient to satisfy assumption $(A5)$, while assumption $(A6)$ imposes some restrictions on the dependence structure among the sub-vectors. If the sub-vectors are mutually independent, then assumption $(A6)$ is clearly satisfied. When the sub-vectors are dependent, additional conditions like weak dependence among the groups of variables are required. In particular, if the sequence $\{\gamma\big({d_i}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2\big), i \geq 1\}$ has the ρ-mixing property, then assumption $(A6)$ holds. A sufficient condition for $\{\gamma\big({d_i}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2\big), i \geq 1\}$ to be a ρ-mixing sequence is to have the sequences $\mathbf{U}$ and $\mathbf{V}$ to satisfy the ρ-mixing property (see Lemma 3 in Appendix A). With these assumptions, we are now ready to state the high-dimensional behavior of $h_b^{\phi,\gamma}$.

Lemma 3.1 *Suppose that $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$ $(1 \leq j, j' \leq J)$ are two independent random vectors satisfying assumptions $(A5)$ and $(A6)$. Additionally, if assumption $(A4)$ is satisfied and ϕ is uniformly continuous, then*

$$\big|h_b(\mathbf{U},\mathbf{V}) - \tilde{h}_b(j,j')\big| \xrightarrow{P} 0 \text{ as } b \to \infty,$$

where $\tilde{h}_b(j,j') \equiv \tilde{h}_b^{\phi,\gamma}(j,j') = \phi\big[b^{-1}\sum_{i=1}^{b} \mathrm{E}\{\gamma(d_i^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2)\}\big]$.

The next result involves $\xi_{jb}(\mathbf{Z})$ (defined in (3.2)) and $\tau_{jb}(\mathbf{Z})$ (defined just above (3.5)), and it is a straightforward extension of Corollary 2.2.

Corollary 3.2 *If a test observation $\mathbf{Z} \sim \mathbf{F}_j$, then for any $1 \leq j' \leq J$, we have*

(a) $\big|\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \tilde{\xi}_b(j,j')\big| \xrightarrow{P} 0$ *as* $b \to \infty$,

(b) $\big|\{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} - \tilde{\tau}_b(j,j')\big| \xrightarrow{P} 0$ *as* $b \to \infty$,

where, for $1 \leq j, j' \leq J$,

$$\tilde{\xi}_b(j,j') \equiv \tilde{\xi}_b^{\phi,\gamma}(j,j') = \tilde{h}_b(j,j') - \frac{1}{2}\big[\tilde{h}_b(j',j') + \tilde{h}_b(j,j)\big], \text{ and}$$

$$\tilde{\tau}_b^{\phi,\gamma}(j,j') \equiv \tilde{\tau}_b(j,j') = \sum_{1\leq l \neq j' \leq J}\left[\frac{n_l}{n-1}|\tilde{h}_b(j',l) - \tilde{h}_b(j,l)|\right] + \frac{n_{j'}-1}{n-1}|\tilde{h}_b(j',j') - \tilde{h}_b(j,j')|.$$

Similar to the constants $\tilde{\xi}_d(j,j')$ and $\tilde{\tau}_d(j,j')$, both $\tilde{\xi}_b(j,j')$ and $\tilde{\tau}_b(j,j')$ are measures of separability between $\mathbf{F}_j$ and $\mathbf{F}_{j'}$ for $1 \leq j, j' \leq J$. While $\tilde{\tau}_b(j,j')$ is non-negative by definition, the same is true for $\tilde{\xi}_b(j,j')$ if ϕ is concave. Moreover, under conditions similar to Lemma 2.3, both $\tilde{\xi}_d(j,j')$ and $\tilde{\tau}_d(j,j')$ are strictly positive whenever $\mathbf{F}_j$ and $\mathbf{F}_{j'}$ have different group distributions (i.e., $\mathbf{F}_{j,i} \neq \mathbf{F}_{j',i}$ for some $1 \leq i \leq b$). This is shown in the following lemma.

Lemma 3.3 *Let γ have non-constant, completely monotone derivative on $\mathbb{R}^+$. Then, the following results hold.*
(a) If ϕ is concave, then $\tilde{\xi}_b(j,j') \geq 0$ for all $1 \leq j, j' \leq J$. Moreover, $\tilde{\xi}_b(j,j') = 0$ if and only if $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$ for all $1 \leq i \leq b$.
(b) If ϕ is one-to-one, then $\tilde{\tau}_b(j,j') = 0$ if and only if $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$ for all $1 \leq i \leq b$.

To derive HDLSS asymptotic results, we require the competing populations to be asymptotically separable. So, we assume the following:

$(A7)$ for every $1 \leq j \neq j' \leq J$, $\liminf_{b \to \infty} \tilde{\xi}_b(j, j') > 0$.

This assumption ensures that separation induced by the blocks is asymptotically non-negligible. It further implies that a similar condition holds for $\tilde{\tau}_b(j, j')$ (see Lemma 1 in Appendix A). Following our discussion preceding Lemma 3.3, assumption $(A7)$ is a generalization of assumption $(A3)$ because if we have difference in the marginal distributions, then the joint distributions are bound to be different. But, the converse is clearly not true. In other words, if two distributions $\mathbf{F}_j$ and $\mathbf{F}_{j'}$ are not separable in terms of $\tilde{\xi}_b$ (respectively, $\tilde{\tau}_b$), then they are not separable in terms of $\tilde{\xi}_d$ (respectively, $\tilde{\tau}_d$). The following theorem shows the high-dimensional behavior of the bgSAVG and NN-bgMADD classifiers under assumption $(A7)$.

Theorem 3.4 *Define $n_0 = \min\{n_1, \ldots, n_J\}$. If assumptions (A4)–(A7) are satisfied, then*
(a) for $n_0 \geq 2$, the misclassification probability of the bgSAVG classifier converges to zero as $b \to \infty$,
(b) for any $k \leq n_0$, the misclassification probability of the k-NN classifier based on bgMADD converges to zero as $b \to \infty$.

Recall that in Examples 2 and 3 we have identical marginal distributions (namely, $N(0, 1)$) for both the classes, but differences in their joint distributions. Theorem 3.4 states that if this information from the joint distributions can be captured by appropriately identifying the groups, then the misclassification probability for both the classifiers should decrease to 0 as d (equivalently, b) increases. We have already observed this in Figure 3.

3.2 Comparison between bgSAVG and NN-bgMADD

In the previous sub-section, we have observed that both bgSAVG and NN-bgMADD classifiers achieve *perfect classification* in high dimensions under similar conditions. But, their relative performance may vary, especially when the dimension is not sufficiently large. To demonstrate the relative behavior of these two classifiers, we now consider two examples. The first example is Example 2 from Section 1. As a second example, we use the following.

Example 4 *We consider two populations, where the d component variables are i.i.d. For the first population, the component distribution is Cauchy with location parameter* 0 *and scale* 1 *(standard Cauchy), while it is Cauchy with location parameter* 0.75 *and scale* 0.75 *for the second one. In this example, we take $n_1 = 50$ and $n_2 = 25$ to form the training set.*

Let us now look into the numerical performance of the proposed classifiers in Examples 2 and 4. We keep all other parameters (e.g., the number of iterations, test sample size) associated with this simulation same as before, and set $d_i = 10$ (respectively, $d_i = 1$) for all $1 \leq i \leq b$ in Example 2 (respectively, Example 4).

Figure 4 clearly shows that the estimated misclassification probabilities for the proposed classifiers (with $\gamma(t) = 1 - e^{-t}$ and $\phi(t) = t$) go to 0 with increasing values of d, and hence quite close to the estimated Bayes risks in Examples 2 and 4. Clearly, assumptions

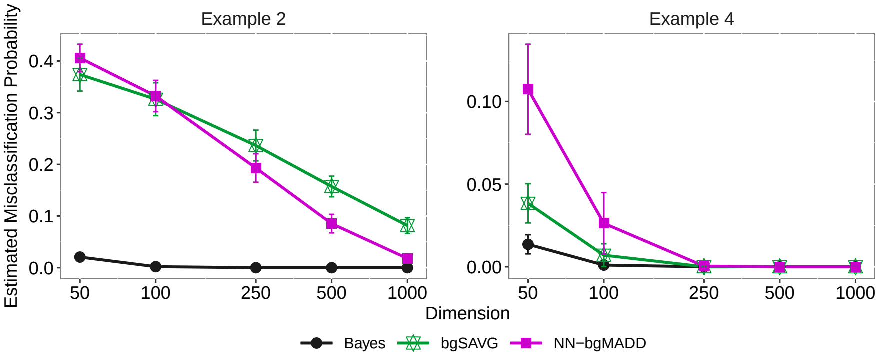

Figure 4: Average misclassification rates (along with the standard errors) based on 100 repetitions for the bgSAVG and NN-bgMADD classifiers are plotted for increasing values of d (in logarithmic scale) for Examples 2 and 4.

$(A4)-(A7)$ hold in both these examples (with bounded γ for Example 4). In Example 2, the block distributions are 10-dimensional multivariate Gaussian with different correlation structures for the two classes. The marginal distributions are Cauchy (i.e., heavy-tailed) in Example 4 with differences in their locations and scales. So, assumptions $(A5)$ and $(A6)$ hold with a bounded γ function. Interestingly, bgSAVG and NN-bgMADD behave differently in these examples with one dominating the other in the respective examples.

Let us now study this phenomena in further detail. From the proof of Theorem 3.4, one can observe that the high-dimensional behavior of the bgSAVG and NN-bgMADD classifiers depend on the behavior of the constants $\tilde{\xi}_b(j,j')$ and $\tilde{\tau}_b(j,j')$, respectively, for $1 \leq j, j' \leq J$. Consequently, the difference between these two classifiers lies in the difference between these constants. To compare between these two classifiers, we make the following assumption, which implies that the difference between $\tilde{\xi}_b(j,j')$ and $\tilde{\tau}_b(j,j')$ does not vanish as the data dimension increases.

$$(A8)\liminf_b |\tilde{\xi}_b(j,j') - \tilde{\tau}_b(j,j')| > 0 \text{ for all } 1 \leq j \neq j' \leq J.$$

The next theorem states the condition under which one classifier dominates the other, and vice-versa. Define the misclassification probabilities as $\Delta_{\text{bgSAVG}} = \mathrm{P}[\delta_{\text{bgSAVG}}(\mathbf{X}) \neq Y]$ and $\Delta_{\text{NN}-\text{bgMADD}} = \mathrm{P}[\delta_{\text{NN}-\text{bgMADD}}(\mathbf{X}) \neq Y]$, where Y denotes the class label of $\mathbf{X}$.

Theorem 3.5 *If assumptions $(A4)-(A6)$ and $(A8)$ are satisfied, and there exists an integer B_1 such that $\tilde{\xi}_b(j,j') > \tilde{\tau}_b(j,j')$ for all $b \geq B_1$ and $1 \leq j \neq j' \leq J$, then there exists an integer B_2 such that*

$$\Delta_{\text{bgSAVG}} \leq \Delta_{\text{NN}-\text{bgMADD}} \textit{ for all } b \geq B_2.$$

Remark 3.6 *If the constants $\tilde{\xi}_b(j,j')$ and $\tilde{\tau}_b(j,j')$ are interchanged in the inequality (stated above), then the ordering of the misclassification probability of the respective classifiers is reversed.*

We now elaborate on this theorem for two-class problems. Recall the expressions for $\tilde{\xi}_b(1,2)$ and $\tilde{\tau}_b(1,2)$ from Corollary 3.2. The ordering between $\tilde{\xi}_b(1,2)$ and $\tilde{\tau}_b(1,2)$ clearly depend on the relationship between the constants $\tilde{h}_b(1,2)$, $\tilde{h}_b(1,1)$ and $\tilde{h}_b(2,2)$ (recall the definition from Lemma 3.1), and the sample sizes n_1 and n_2. A detailed case by case study on this inequality is provided by Lemma 2 in Appendix A. To draw a comparison, let us now look back at Examples 2 and 4. Clearly, the constants $\tilde{\xi}_b(1,2)$, $\tilde{\tau}_b(1,2)$ and $\tilde{\tau}_b(2,1)$ are free of b in both these examples. Calculating the constants involve computing univariate/multivariate integrals. More details on these calculations can be found in Section 2 of the Supplementary. The constants take the values $\tilde{\xi}_b(1,2) = 0.0101$, $\tilde{\tau}_b(1,2) = 0.0470$ and $\tilde{\tau}_b(2,1) = 0.0472$ in Example 2, while in Example 4 they are $\tilde{\xi}_b(1,2) = 0.0327$, $\tilde{\tau}_b(1,2) = 0.0213$ and $\tilde{\tau}_b(2,1) = 0.0222$ (also see Table 1). Clearly, the value of $\tilde{\xi}_b(1,2)$ is smaller than those of $\tilde{\tau}_b(1,2)$ and $\tilde{\tau}_b(2,1)$ in Example 2. Theorem 3.5 suggests that the misclassification probability of the NN-bgMADD classifier should be smaller than the bgSAVG classifier for large values of b. This can be observed in the left panel of Figure 4 for dimension higher than 100. On the other hand, in Example 4, the value of $\tilde{\xi}_b(1,2)$ is larger than those of $\tilde{\tau}_b(1,2)$ and $\tilde{\tau}_b(2,1)$, and one observes a role reversal in the right panel of Figure 4. This analysis has been continued for all the examples discussed in this article later in Section 5.

A few words are called for assumption ($A8$), which holds under various scenarios. In particular, if the component variables of the underlying distributions are i.i.d., then $\tilde{\xi}_b$ and $\tilde{\tau}_b$ are free of b. Some more general conditions are discussed in Lemma 2 of Appendix A. It can also be shown that assumption ($A8$) holds under more general cases like Example 2 (see Remark A in Appendix A).

4. Practical Implementation of Variable Clustering

For practical implementation of the methodology defined in the previous section, we need to find an appropriate clustering $\mathcal{C}$ of the component variables. The basic idea is to partition a d-dimensional vector $\mathbf{U}$ into b disjoint groups (or, sub-vectors) $\mathbf{U}_1, \ldots, \mathbf{U}_b$ such that the variables in the same sub-vector are more *similar* to each other than the variables in different sub-vectors. Such phenomena (groups of variables) arises naturally in scientific areas like genomics. In microarray gene expressions, genes that share similar pattern of expression are usually put into a cluster (see, e.g., Eisen et al., 1998), while such groups of variables also play a key role in bio-diversity modeling (see, e.g., Faith and Walker, 1996).

We would like to emphasize that the order in which the component variables are arranged in a sub-vector is irrelevant in this context. Therefore, we use the terms 'group' and 'sub-vector' interchangeably. Here, we assume the same grouping of component variables for all J populations. In general, different populations may have different groups of component variables. But, in a two-class problem, if the group structure of one population is either finer (or, coarser) w.r.t. the other population, then we can assume the coarser structure for both the populations. For more than two classes, if the group structure of one population is coarser than all the competing populations, it is sufficient to use the coarsest structure across all populations. In any case, our problem is essentially that of clustering d variables with n observations for each variable (i.e., d observations in $\mathbb{R}^n$). Any appropriate clustering algorithm (see, e.g., Hastie et al., 2009) can be used for this purpose. To summarize, one can view this idea of constructing groups as a problem of clustering the component variables using an appropriate measure of similarity. So first, let us discuss the idea of *similarity* (equivalently, *dissimilarity*) among variables.

For the HDLSS asymptotic results, we need variables from different groups (or, clusters) to have weak dependence (see assumption $(A6)$). On the other hand, highly dependent variables are natural candidates to be included in the same cluster. A reasonable measure of dependence between two components is the absolute value of their correlation coefficient. Let $r(i, i')$ denote the correlation between the i-th and the i'-th components for $1 \leq i, i' \leq d$. If $|r(i, i')|$ is high, then we say that the i-th and the i'-th components are strongly associated, or 'similar'. While $|r(i, i')|$ is a measure of similarity, $1 - |r(i, i')|$ can be considered as a measure of *dissimilarity*. We use the agglomerative hierarchical clustering algorithm with average linkage (see, e.g., Hastie et al., 2009) and $1 - |r(i, i')|$ as the pairwise dissimilarity measure to obtain clusters of components. Starting with each component variable as a single cluster, hierarchical methods merge the least dissimilar clusters in turn until all the components are put together in one single cluster. For heavy-tailed distributions (like the Cauchy distribution), a robust measure of correlation can be used.

In hierarchical clustering, each level in the hierarchy induces a set of clusters, and the whole hierarchy (visualized as a dendrogram) represents a nested structure among the clusters obtained at different levels (see Figure 5 below). The height of each level represents the dissimilarity between the clusters that are merged together at that level. In other words, each cluster structure is represented by the height of the level corresponding to that structure. Therefore, finding an appropriate clustering is equivalent to identifying a suitable level in the hierarchy. Suppose $\mathbf{H}$ is the set of all heights that are obtained at different levels of clustering. We order the values in $\mathbf{H}$, and find the α-th percentile H_α for different values of $\alpha \in A = \{0, 0.1, \ldots, 0.9, 1\}$. For each fixed α, we obtain a clustering induced by H_α. Note that the number of clusters is non-increasing in α, while the size of each cluster is non-decreasing. In particular, H_0 corresponds to the case where each cluster consists of a single component variable only, i.e., $b = d$. On the other hand, H_1 leads to the clustering where all the d components are put together in a single cluster.

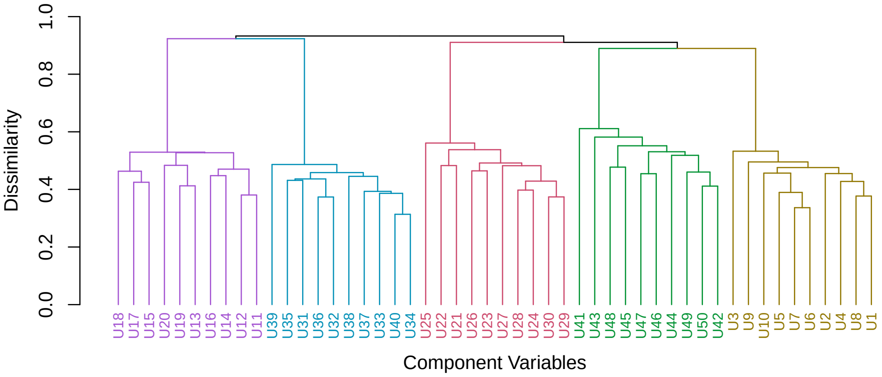

Figure 5: Dendrogram showing structures of clusters in Example 2 for one run of a simulation with $d = 50$.

We demonstrate this idea using Example 2. In this example (with $d = 50$ and $n_1 = n_2 = 50$), the groups of component variables (common across both classes) are the sets $C_1 = \{1, \ldots, 10\}; C_2 = \{11, \ldots, 20\}; \ldots; C_5 = \{41, \ldots, 50\}$. We consider a simulated realization from this example. Figure 5 shows the dendrogram for this data. At $H_{0.9} = 0.67$, we obtain five clusters in Figure 5. The distinct clusters are indicated with five different colors, while the components corresponding to each cluster are marked with the same color in Figure 5. Clearly, the method correctly assigns desired components to the respective groups (up to a permutation of the components within each group). Once the groups $\mathbf{U}_1, \ldots, \mathbf{U}_b$ have been identified, we can compute $h_b^{\phi,\gamma}$ as in equation (3.1) and classify observations using the bgSAVG classifier, or the NN-bgMADD classifier introduced in Section 3.

It is evident from Figure 5 that the choice of H_α (or, equivalently α) is crucial in finding the 'true' cluster structure. However, our task here is not to find the 'true' cluster structure in the variables, but rather to find cluster structures that are useful for classification. Similar to the cluster structure, the performance of a classifier should also depend on the choice of α. To investigate this, we looked at the misclassification rates of the bgSAVG and the NN-bgMADD classifiers (with $\gamma(t) = 1 - e^{-t}$ and $\phi(t) = t$) in Examples 2–4 for varying choices of α (which corresponds to different cluster structures). Clearly, Figure 6 shows that the classification performance depends crucially on the choice of α.

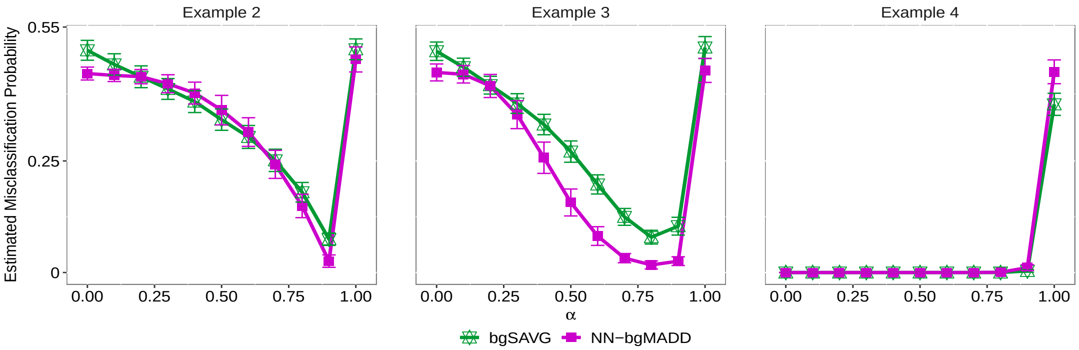

Figure 6: Average misclassification rates (along with the standard errors) based on 100 repetitions of the bgSAVG and NN-bgMADD classifiers for increasing values of α in Examples 2–4.

To obtain a data driven choice of α, we use the idea of leave-one-out cross-validation method (see, e.g., Hastie et al., 2009). For a fixed value of $\alpha \in A$, define

$$e_\alpha = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\delta_\alpha^{-i}(\mathbf{X}_i) \neq Y_i\}.$$

Here, δ_α^{-i} is a classifier (bgSAVG or NN-bgMADD) constructed by leaving out the i-th sample from the training data for $1 \leq i \leq n$. Define $\hat{\alpha} = \arg\min_{\alpha \in A} e_\alpha$. We use the clustering induced by $H_{\hat{\alpha}}$ as the optimal one to carry out further analysis.

As we already mentioned, the idea of grouping in component variables can be found in several real data scenarios as well. To realize this, we plot similarity matrices of the components for `four` high-dimensional data sets from `three` different data archives. The `Cricket X` and `EOGHorizontalSignal` data sets are both 12 class problems from the UCR Time Series Classification Archive (see Dau et al., 2018) with (n_j, d) as $(32, 300)$ and $(30, 1250)$ for $1 \leq j \leq J$. The first data is related to motion, while the second data set was collected from an electro-oculography (EOG). In Figure 7, we distinctly observe about 1 group and 2 groups (the second group has some smaller blocks) for these two data sets, respectively. The GSE2685 data set (available at the Microarray database: http://www.biolab.si/supp/bi-cancer/projections/) comprises of gene expression measurements of 30 tissue samples distributed over 3 classes (8 normal gastric tissues, 5 diffuse gastric tumors and 17 intestinal gastric tumors). The blocks are unclear if we plot all 4522 genes (variables) in this data set, so we have created a plot with reduced number of (about 1500) variables. In the nutt2003v2 data set (available at the Compcancer database: https://schlieplab.org/Static/Supplements/CompCancer/datasets.htm), it was investigated whether gene expression profiling could be used to classify high-grade gliomas. Microarray analysis was used to determine the expression of approximately 12000 genes in a set of 28 glioblastomas which were classified as classic (C), or non-classic (N). The plots in Figure 7 also indicate the presence of group structure in these two gene expression data sets. We give a more detailed analysis of these four real data sets later in Section 6.

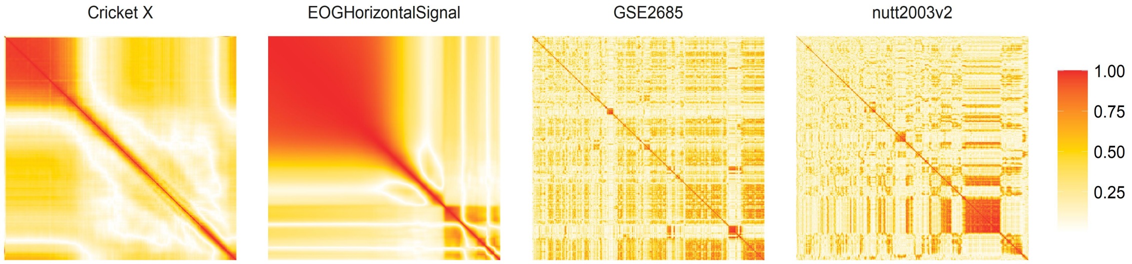

Figure 7: Absolute of sample correlation matrices for the four benchmark data sets.

5. Simulation Studies

In this section, we thoroughly analyze some high-dimensional simulated data sets to compare the performance of the classifiers proposed in Sections 2 and 3. We have already introduced Examples 1–3 in Section 1, and Example 4 in Section 3. Four new examples are considered in this section to demonstrate the performance of the proposed classifiers.

Example 5 *The two distributions are $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(0.25\mathbf{1}_d, \mathbf{I}_d)$, where $\mathbf{0}_d$ is the d-dimensional vector of zeros, $\mathbf{1}_d$ is the d-dimensional vector of ones and $\mathbf{I}_d$ is the $d \times d$ identity matrix. Note that the d component variables are i.i.d. for both the populations.*

Example 6 *We again consider two Gaussian distributions $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\mathbf{0}_d, 0.5\mathbf{I}_d)$. Here, the d component variables are i.i.d. similar to Example 5.*

Example 7 *The distributions are $\mathbf{F}_1(\mathbf{u}) = \prod_{i=1}^{2} \mathbf{F}_{1,i}(\mathbf{u}_i)$ and $\mathbf{F}_2(\mathbf{u}) = \prod_{i=1}^{2} \mathbf{F}_{2,i}(\mathbf{u}_i)$, with $\mathbf{F}_{1,1} \equiv N_{\lfloor \frac{d}{2} \rfloor}(\mathbf{0}_{\lfloor \frac{d}{2} \rfloor}, \mathbf{I}_{\lfloor \frac{d}{2} \rfloor})$, $\mathbf{F}_{1,2} \equiv N_{d-\lfloor \frac{d}{2} \rfloor}(\mathbf{0}_{d-\lfloor \frac{d}{2} \rfloor}, 0.5\mathbf{I}_{d-\lfloor \frac{d}{2} \rfloor})$, $\mathbf{F}_{2,1} \equiv N_{\lfloor \frac{d}{2} \rfloor}(\mathbf{0}_{\lfloor \frac{d}{2} \rfloor}, 0.5\mathbf{I}_{\lfloor \frac{d}{2} \rfloor})$ and $\mathbf{F}_{2,2} \equiv N_{d-\lfloor \frac{d}{2} \rfloor}(\mathbf{0}_{d-\lfloor \frac{d}{2} \rfloor}, \mathbf{I}_{d-\lfloor \frac{d}{2} \rfloor})$. Here, $\lfloor \cdot \rfloor$ denotes the floor function.*

Example 8 *We take $\mathbf{F}_1 \equiv N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $\mathbf{F}_2(\mathbf{u}) = \prod_{i=1}^{b} \mathbf{F}_{2,i}(\mathbf{u}_i)$, with $\mathbf{F}_{2,i} \equiv PN_{10}(\mathbf{1}_{10}, 10)$ for all $1 \leq i \leq b$. Here, $PN_{10}(\boldsymbol{\beta}, \alpha)$ denotes the ten-dimensional multivariate power normal distribution with parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})^\top$ with $\beta_i > 0$ for all $1 \leq i \leq 10$ and $\alpha > 0$*

(see, e.g., Kundu and Gupta, 2013). Note that $\beta_i = 1$ for all $1 \leq i \leq 10$ implies that the one-dimensional marginals of $\mathbf{F}_2$ are all standard normal.

In each example, we simulated data for $d = 50$, 100, 250, 500 and 1000. The training sample was formed by generating 50 observations from each class (except Example 4) and a test set of size 500 (250 from each class) was used. In Example 4, the training samples sizes were set to be 50 and 25, respectively. This process was repeated 100 times to compute the average misclassification rates, which are reported in Figure 8. For the proposed generalized and block-generalized classifiers, we used $\gamma(t) = 1 - e^{-t}$ and $\phi(t) = t$.

Observe that in Examples 1, 2, 3, 7 and 8, we have $\boldsymbol{\mu}_{1d} = \boldsymbol{\mu}_{2d} = \mathbf{0}_d$ (i.e., $\nu_{12}^2 = 0$). Furthermore, we have $\sigma_1^2 = \sigma_2^2 = 5/3$ in Example 1 and $\sigma_1^2 = \sigma_2^2 = 0.75$ in Example 7, while $\sigma_1^2 = \sigma_2^2 = 1$ in Examples 2, 3 and 8. This implies that $\sigma_1^2 - \sigma_2^2 = 0$ for all these five examples. In Example 4, the moment based quantities ν_{12}^2, σ_1^2 and σ_2^2 do not exist as the underlying distributions are Cauchy. On the other hand, Example 5 is a *location* problem ($\nu_{12}^2 = 0.25$ with $\sigma_1^2 - \sigma_2^2 = 0$), while Example 6 is a *scale* problem ($\nu_{12}^2 = 0$ with $|\sigma_1^2 - \sigma_2^2| = 0.5$). In our earlier analysis of Examples 1–4, we assumed the group information $\mathcal{C}$ to be *known*. We now analyze all eight examples to validate the fact that the data driven procedure for blocking the variables (developed in Section 4) in combination with the block-generalized classifiers (proposed in Section 3) yield promising performance in high dimensions.

In Examples 1, 4, 5, 6 and 7, the component variables are i.i.d. and the populations have differences in their one-dimensional marginals. So, assumptions $(A1) - (A3)$ are satisfied and consequently, the misclassification probabilities of the gSAVG and NN-gMADD classifiers are close to zero (see Figure 8). This is not the case for the other three examples. In Examples 2, 3 and 8, the one-dimensional marginals are standard normal for both populations, so assumption $(A3)$ is clearly violated. We observe that both the gSAVG and NN-gMADD classifiers misclassify nearly half of the test points in these examples. On the other hand, assumptions $(A5) - (A7)$ are satisfied for these examples. So, the bgSAVG and NN-bgMADD classifiers classify almost all the test points correctly. Blocks of variables were estimated using the method described in Section 4, where we used the absolute value of Pearson's correlation coefficient as the measure of similarity. However, this measure is inappropriate for Example 4 (with Cauchy distributions). So, we have used the minimum regularized covariance determinant (MCD) estimator, which is available through the `R` package `rrcov`. We observe that the estimated misclassification probabilities of the bgSAVG and NN-bgMADD classifiers are very close to zero in high dimensions (see Figure 8), which is consistent with the idea of *perfect classification* as $b \to \infty$ (also see Theorem 3.4).

A question that arises naturally from Figure 8 is the relative performance of the bgSAVG classifier and the NN-bgMADD classifier for moderate values of d. In Section 3.2, we used Examples 2 and 4 to motivate this question and investigated this fact theoretically in Theorem 3.5. We now complete this investigation for the other examples. Recall that the relative performance of these two classifiers depends on the ordering of the constants $\tilde{\xi}_b(1,2)$, and $\tilde{\tau}_b(1,2), \tilde{\tau}_b(2,1)$ (see Theorem 3.5 and the preceeding discussion). We have computed the value of these constants in Table 1. Section 2 of the Supplementary contains more details and related calculations.

We can observe from Figure 8 that the NN-bgMADD classifier performs better than the bgSAVG classifier in Examples 1, 2, 3 and 6 for moderate values of d ($\sim 100 - 250$).

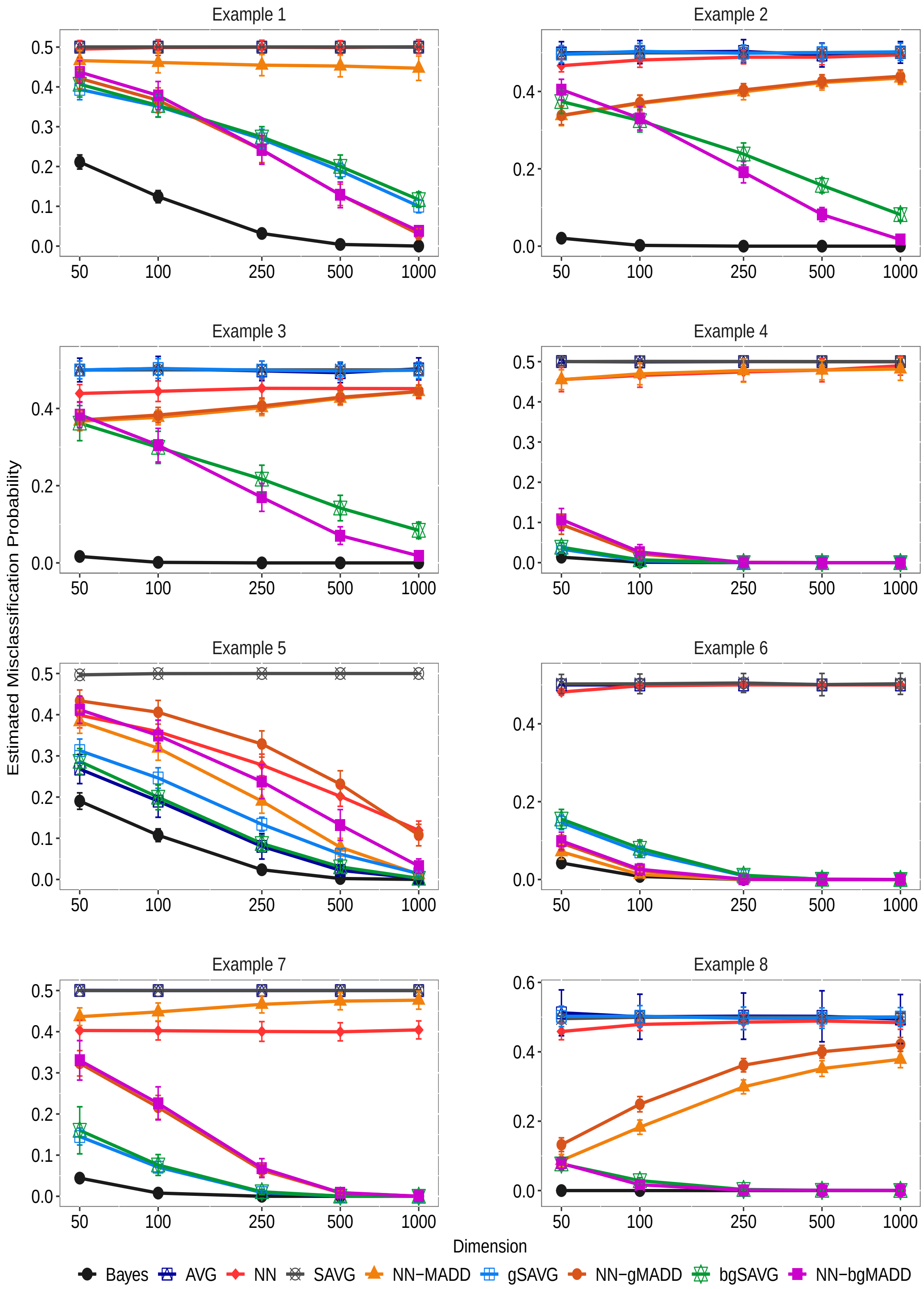

Figure 8: Average misclassification rates (along with the standard errors) based on 100 repetitions for different classifiers are plotted for increasing values of d (in logarithmic scale).

On the contrary, the bgSAVG classifier clearly dominates the NN-bgMADD classifier in Examples 4, 5, 7 and 8. This phenomena is consistent with the ordering of $\tilde{\xi}_b(1,2)$, and $\tilde{\tau}_b(1,2), \tilde{\tau}_b(2,1)$ in Table 1, except in Examples 5 and 7, where the value of these constants are equal. Interestingly, the bgSAVG classifier performs better than the NN-bgMADD classifier in these two examples. This can be explained by looking closer into the expression of these constants. Recall from Corollary 3.2 that these constants involve the terms $\tilde{h}_b(1,1)$, $\tilde{h}_b(2,2)$ and $\tilde{h}_b(1,2)$. The fact that $\tilde{h}_b(1,2) > \max\{\tilde{h}_b(1,1), \tilde{h}_b(2,2)\}$ (see the values for Examples 5 and 7 in Table 1) justifies the improved performance of the bgSAVG classifier (also see Sarkar et al. (2020) for related explanations in the context of two sample testing).

Table 1: Values of the constants $\tilde{\xi}_b(1,2)$, $\tilde{\tau}_b(1,2)$ and $\tilde{\tau}_b(2,1)$ in Examples 1–8. The figure in **bold** indicates the maximum of these three values.

Ex.	$\tilde{h}_b(1,1)$	$\tilde{h}_b(2,2)$	$\tilde{h}_b(1,2)$	$\tilde{\xi}_b(1,2)$	$\tilde{\tau}_b(1,2)$	$\tilde{\tau}_b(2,1)$
1	0.6387	0.6017	0.6230	0.0027	**0.0185**	**0.0185**
2	0.7909	0.6967	0.7539	0.0101	0.0470	**0.0472**
3*	0.7614	0.7091	0.7423	0.0070	0.0260	**0.0262**
4	0.7440	0.6789	0.7442	**0.0327**	0.0213	0.0222
5	0.5528	0.5528	0.5583	**0.0056**	**0.0056**	**0.0056**
6	0.5528	0.4226	0.5000	0.0123	0.0649	**0.0652**
7	0.4877	0.4877	0.5000	**0.0123**	**0.0123**	**0.0123**
8	0.8138	0.5903	0.7634	0.0614	0.1111	**0.1124**

* the block size (d_i) was fixed at 5

5.1 Comparison with popular classifiers

Here, we compare the performance of the proposed classifiers with some well-known classifiers, namely, Support Vector Machines (SVM, Vapnik, 1998), GLMNET (Hastie et al., 2009), neural networks (NNET, Bishop, 1995) and nearest neighbor classifiers based on the random projection method (NN-RAND, Deegalla and Bostrom, 2006). We studied numerical performance of these classifiers for $d = 1000$ (see Tables 2 and 3 in the Supplementary for other values of d). The average misclassification rates along with the corresponding standard errors are reported in Table 2. Misclassification rates of both the linear and non-linear SVM are reported. We used the radial basis function (RBF) kernel, i.e., $K_\theta(\mathbf{x}, \mathbf{y}) = \exp\{-\theta\|\mathbf{x}-\mathbf{y}\|^2\}$ in non-linear SVM with $\theta \in \{i/10d;\ 1 \leq i \leq 20\}$ and reported the minimum misclassification rate. For NNET, we used the sigmoid as its activation function. The number of hidden layers were allowed to vary in the set $\{1, 3, 5, 10\}$, and the minimum misclassification rate was reported as NNET. We have used default values for the other parameters that were involved with these classifiers. The `R` packages `e1071`, `glmnet`, `RSNNS` and `RandPro` were used for SVM, GLMNET, NNET and NN-RAND, respectively. Our classifiers were implemented in `R` too, and the codes are available from this link. We fix $\phi(t) = t$ for the proposed classifiers. Untill this point, we have used the choice $\gamma_1(t) = 1-e^{-t}$ only. We now introduce two more choices of γ, namely, $\gamma_2(t) = \log(1+t)$ and $\gamma_3(t) = \sqrt{t}/2$ in this section. For our proposed methods, we report the misclassification rates for all three choices of γ in Table 2.

Table 2: Misclassification rates (stated in the first row) and standard errors (stated in the second row) of different classifiers in Examples 1–8 for $d = 1000$. The figure in **bold** indicates the minimum misclassification rate.

Ex.	GLMNET	NN	SVM	SVM	NNET	gSAVG			bgSAVG			NN-gMADD			NN-bgMADD		
		-RAND	-LIN	-RBF		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
1	0.4748	0.4972	0.4979	0.4952	0.4919	0.1002	0.2079	0.2646	0.1167	0.2156	0.2702	**0.0302**	0.1321	0.2451	0.0379	0.1411	0.2457
	0.0177	0.0171	0.0232	0.0203	0.0240	0.0194	0.0195	0.0208	0.0165	0.0229	0.0230	0.0102	0.0260	0.0314	0.0135	0.0274	0.0374
2	0.4745	0.4940	0.5099	0.4540	0.5010	0.5025	0.5029	0.5024	0.0815	0.1243	0.1461	0.4445	0.4390	0.4384	0.0185	0.0171	**0.0168**
	0.0174	0.0150	0.0208	0.0226	0.0253	0.0223	0.0228	0.0224	0.0152	0.0201	0.0208	0.0166	0.0174	0.0173	0.0088	0.0084	0.0084
3	0.4757	0.4558	0.5000	0.5000	0.4997	0.4991	0.5011	0.5018	0.0843	0.1431	0.1532	0.4495	0.4442	0.4443	0.0185	0.0184	**0.0182**
	0.0182	0.0279	0.0000	0.0000	0.0232	0.0214	0.0230	0.0227	0.0214	0.0260	0.0269	0.0165	0.0152	0.0161	0.0100	0.0105	0.0105
4	0.4173	0.4933	0.4282	0.4995	0.3688	**0.0000**	**0.0000**	0.0017	**0.0000**	**0.0000**	0.0022	**0.0000**	0.0007	0.2319	**0.0000**	0.0009	0.2279
	0.0266	0.0245	0.0205	0.0014	0.0236	0.0000	0.0000	0.0018	0.0000	0.0000	0.0021	0.0000	0.0016	0.0341	0.0000	0.0018	0.0313
5	0.2172	0.0336	0.0018	0.0012	0.2748	0.0142	0.0022	0.0018	0.0028	**0.0007**	**0.0007**	0.1078	0.0248	0.0202	0.0325	0.0139	0.0134
	0.0220	0.0139	0.0020	0.0017	0.0444	0.0055	0.0020	0.0017	0.0028	0.0014	0.0014	0.0261	0.0102	0.0092	0.0173	0.0088	0.0089
6	0.4533	0.5000	0.4587	0.0000	0.4968	**0.0000**	**0.0000**	0.0003	**0.0000**	**0.0000**	0.0003	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	0.0158	0.0000	0.0153	0.0000	0.0238	0.0000	0.0000	0.0009	0.0000	0.0003	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.4677	0.3977	0.4974	0.4694	0.4968	**0.0000**	**0.0000**	0.0002	**0.0000**	**0.0000**	0.0002	0.0001	0.0034	0.0143	0.0001	0.0034	0.0148
	0.0184	0.0245	0.0240	0.0228	0.0218	0.0000	0.0002	0.0006	0.0000	0.0000	0.0005	0.0005	0.0036	0.0067	0.0004	0.0034	0.0066
8	0.4767	0.5000	0.5010	0.2106	0.4971	0.5001	0.4987	0.4969	**0.0003**	0.0028	0.0033	0.4036	0.3914	0.3883	0.0005	0.0022	0.0024
	0.0153	0.0233	0.0208	0.0218	0.0231	0.0273	0.0328	0.0328	0.0013	0.0050	0.0064	0.0218	0.0245	0.0240	0.0015	0.0042	0.0048

Table 3: Misclassification rates (stated in the first row) and standard errors (stated in the second row) of different classifiers in four benchmark data sets. The figure in **bold** indicates the minimum misclassification rate.

Data	GLMNET	NN	SVM	SVM	NNET	gSAVG			bgSAVG			NN-gMADD			NN-bgMADD		
		-RAND	-LIN	-RBF		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
CricketX	0.6553	0.5039	0.6061	0.4154	0.6643	0.6513	0.6500	0.6472	0.6008	0.6215	0.6167	0.3756	0.3907	0.3929	**0.3326**	0.3612	0.3660
	0.0184	0.0228	0.0212	0.0210	0.0263	0.0201	0.0231	0.0220	0.0279	0.0233	0.0250	0.0218	0.0207	0.0211	0.0212	0.0210	0.0222
EOGHorizontal Signal	0.4824	0.4141	0.4691	0.4241	0.7280	0.7334	0.5379	0.5028	0.7135	0.4673	0.4684	0.8524	0.5048	0.4998	0.8788	**0.2938**	0.3475
	0.0183	0.0241	0.0236	0.0211	0.0458	0.0183	0.0231	0.0201	0.0127	0.0236	0.0236	0.0170	0.0214	0.0254	0.0153	0.0205	0.0181
GSE2685	0.2060	0.2913	**0.1787**	0.3475	0.4013	0.5213	0.4781	0.4763	0.4438	0.4263	0.4175	0.3575	0.2869	0.2381	0.4480	0.2120	0.2873
	0.0622	0.1091	0.0613	0.0505	0.1081	0.1159	0.1282	0.1252	0.1413	0.1370	0.1442	0.0875	0.0941	0.0887	0.1396	0.0959	0.1104
nutt2003v2	0.1993	0.4000	0.1114	0.2100	0.4993	0.3336	0.2150	0.1871	0.3514	0.0871	**0.0779**	0.3686	0.1957	0.1557	0.2593	0.1286	0.1186
	0.1081	0.0825	0.0769	0.1695	0.0864	0.1264	0.1082	0.1102	0.1039	0.0588	0.0509	0.0951	0.0784	0.0762	0.1229	0.0626	0.0549

In all the examples (except Example 5), the competing classifiers GLMNET, NN-RAND, SVM and NNET misclassify almost 50% of the test sample points. Example 5 involves a *location* problem, and all these popular classifiers perform quite well, with SVM having a clear edge over the others, followed closely by NN-RAND. The non-linear classifier SVM-RBF leads to *perfect classification* in Example 6 (a *scale* problem), and an improved misclassification rate of about 21% in Example 8 (having differences in their scatter matrices).

To summarize the performance of our classifiers in Table 2, we observe that the proposed bgSAVG and NN-bgMADD classifiers outperform popular classifiers in all examples. In Example 1, the misclassification rates of these classifiers are slightly more than those of the gSAVG and NN-gMADD classifiers, respectively. We have difference in marginal distributions, and it is not necessary to use variable clustering in this example. The same is true for Examples 4 and 7 as well, but the misclassification rates of the bgSAVG and NN-bgMADD classifiers are quite similar to those of the gSAVG and NN-gMADD classifiers in these two examples. In fact, the additional error incurred due to estimation of groups is negligible in such cases. Moreover, the block-generalized classifiers improve over the generalized classifiers in Example 5. These examples clearly show that block-generalized classifiers perform well even when it is not necessary to group the component variables.

5.2 Comparison among the choices of γ

A natural question that arises from Table 2 is the choice of γ in practice. We have considered three choices of γ, namely, $\gamma_1(t) = 1 - e^{-t}$, $\gamma_2(t) = \log(1+t)$ and $\gamma_3(t) = \sqrt{t}/2$. All these functions have non-constant, completely monotone derivatives (see, e.g., Feller, 1971; Baringhaus and Franz, 2010). These functions are monotonically increasing and there exists a $C > 0$ such that these functions satisfy the ordering $\gamma_1(t) < \gamma_2(t) < \gamma_3(t)$ for all $t > C$. The function γ_1 is clearly bounded, while the other two functions are unbounded. For large t, the function γ_2, although unbounded, stays closer to γ_1 when compared with the function γ_3. The main idea behind choosing these functions was to explore the complete spectrum (i.e., bounded, unbounded and in-between), and understand the effectiveness of the choice of the γ function in capturing discriminative information from the two class distributions.

We deal with heavy-tailed distributions in Example 4, and the advantage of using a bounded γ is clear here. In this example, generalized classifiers based on γ_1 outperformed those based on γ_3. The performance of classifiers based on γ_2 was quite close to γ_1. The fact that γ_1 is a bounded function is necessary here to ensure that assumptions $(A1)$ and $(A2)$ hold. In Example 5 (a *location* problem) involving light-tailed distributions, generalized classifiers based on γ_3 clearly outperform those constructed using γ_1, while the performance of γ_2 again lies in-between these two choices. A related phenomena was also observed by Baringhaus and Franz (2010) for location problems, where the authors were interested in non-parametric two sample goodness of fit tests in $\mathbb{R}^d$. Observe that if we fix a classifier (say, bgSAVG) in Table 2, then either γ_1 (in Examples 1–4 and 6–8) or γ_3 (in Example 5) leads to the minimum misclassification rate. From the results of our simulation study in Table 2, there is *no clear winner* among these two choices of the γ function. So, we recommend using both choices, namely, γ_1 and γ_3 to obtain a complete picture of the underlying scenario.

6. Real Data Analysis

Now, we study the performance of our proposed classifiers on other benchmark data sets from three popular databases, namely, Compcancer database, Microarray database and UCR Time Series Archive (2018). Detailed description of the data sets are available at the respective sources. Data sets in the Compcancer and Microarray databases (involving gene expression studies) have a *fixed data* with corresponding class labels, while those from the UCR Archive come in two parts, a *fixed training set* as well as a *fixed test set*. For our analysis of the data sets in the Compcancer and Microarray databases, we randomly selected 50% of the observations (without replacement) corresponding to each class to form the training set. The rest of the observations were considered as test cases. For data sets from the UCR Archive, 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 over different splits of the data set to obtain a stable estimate of the misclassification rate.

Let us start by analyzing the four benchmark data sets mentioned in Section 4. The numerical results are reported in Table 3. The NN-bgMADD classifier captures information from the group structure and leads to the minimum overall misclassification rate in both `Cricket X` and `EOGHorizontalSignal` data sets. In the `EOGHorizontalSignal` data, we observed a significant variability in the misclassification rates for different choices of γ. In fact, γ_1 (a bounded function) led to a misclassification rate of about 88%. This deteriorating performance of γ_1 may be attributed to the fact that this function involves the term e^{-t}, which reduces the large differences in componentwise means of the competing classes, while γ_3 involves the term $\sqrt{t}/2$, and manages to retain this information. The next two data sets are related to gene expression studies, and the component variables often have differences in their class means. SVM-LIN yields the lowest misclassification rate, while the NN-bgMADD classifier had the second best performance in the `GSE2685` data set. The bgSAVG classifier leads to the best performance in the high-dimensional `nutt2003v2` data, followed by the SVM-LIN and NN-bgMADD classifiers. Generally, we observe that block-generalized classifiers perform significantly better than their generalized counterparts in all four data sets. This further establishes the usefulness of such classifiers in real data scenarios.

The Compcancer database has 35 data sets, while the Microarray database consists of 20 data sets. We chose data sets with $\min_j n_j \geq 6$, which left us with 31 data sets from the first database, and 20 data sets in the second database. The `ALLGSE412` data set in the Microarray database has missing values in 29 observations (out of the 55 samples) corresponding to 14 covariates, so we dropped those covariates from all the samples during our analysis. We used 71 (out of available 85) data sets from the UCR data base.

To begin with, we look at the performance of the generalized and block-generalized classifiers w.r.t. their classical counterparts. In Figure 9, we show boxplots of the misclassification rates for the proposed classifiers, separately for the three databases. It is clear from these figures that the generalized versions of the AVG classifier yield substantial improvement over the usual classifiers, while the block-generalized classifiers yield further improvement in all three databases. However, this improvement is not so compelling for the generalized and block-generalized NN classifiers. Interestingly, *simple* classifiers like

SAVG and NN yield competitive performance in the first two databases involving gene expression studies.

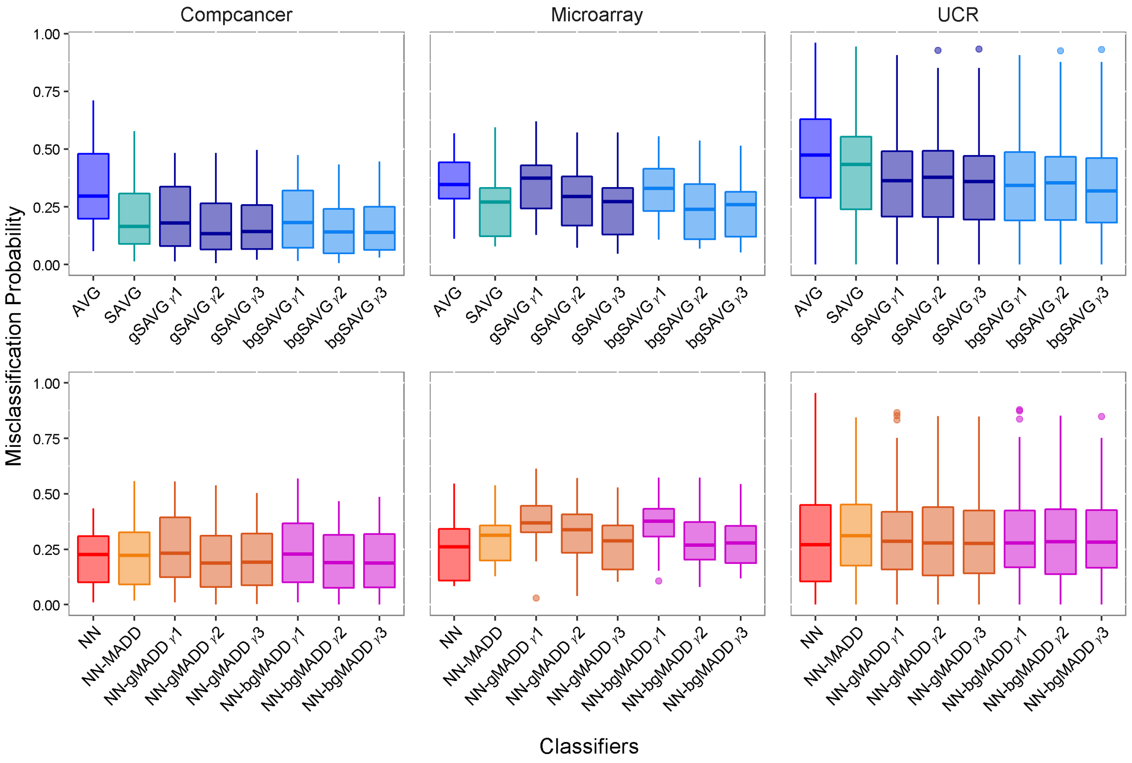

Figure 9: Boxplot of the estimated misclassification probabilities corresponding to various AVG and NN classifiers in the Compcancer, Microarray and UCR databases.

Next, we compare the performance of our proposed classifiers with some existing classifiers (namely, SVM, GLMNET, NNET and NN-RAND). To get an overall picture of their performance in the three databases, we summarized the entire information through boxplots in Figure 10 separately for these three databases. For each database, we considered a boxplot of misclassification rates for all 22 classifiers across all data sets in that database. Detailed results are available in Section 5 (see Tables 4–11) of the Supplementary.

The Compcancer and Microarray databases have datasets involving gene expressions, which are very high-dimensional ($d \sim 1400 - 23000$) with low sample sizes ($n \sim 10 - 100$). Most of these data sets involve 2 or 3 class problems. Linear SVM performs best in these two databases (see Figure 10) since the competing classes often have differences in their mean vectors. GLMNET (a *regularized* linear classifier) induces drastic reduction in the data dimension (the reduced dimension $\sim 1 - 99$), and takes the second position. These data sets have sparsity in their components, which justifies the good performance of GLMNET. However, blocks of variables contain important information (recall panels (c) and (d) of Figure 7) and also lead to dimension reduction through the estimated block structure. This helps the bgSAVG classifier to perform quite well too in these two data bases. Generally, the bgSAVG classifier tends to perform better than the NN-bgMADD classifier.

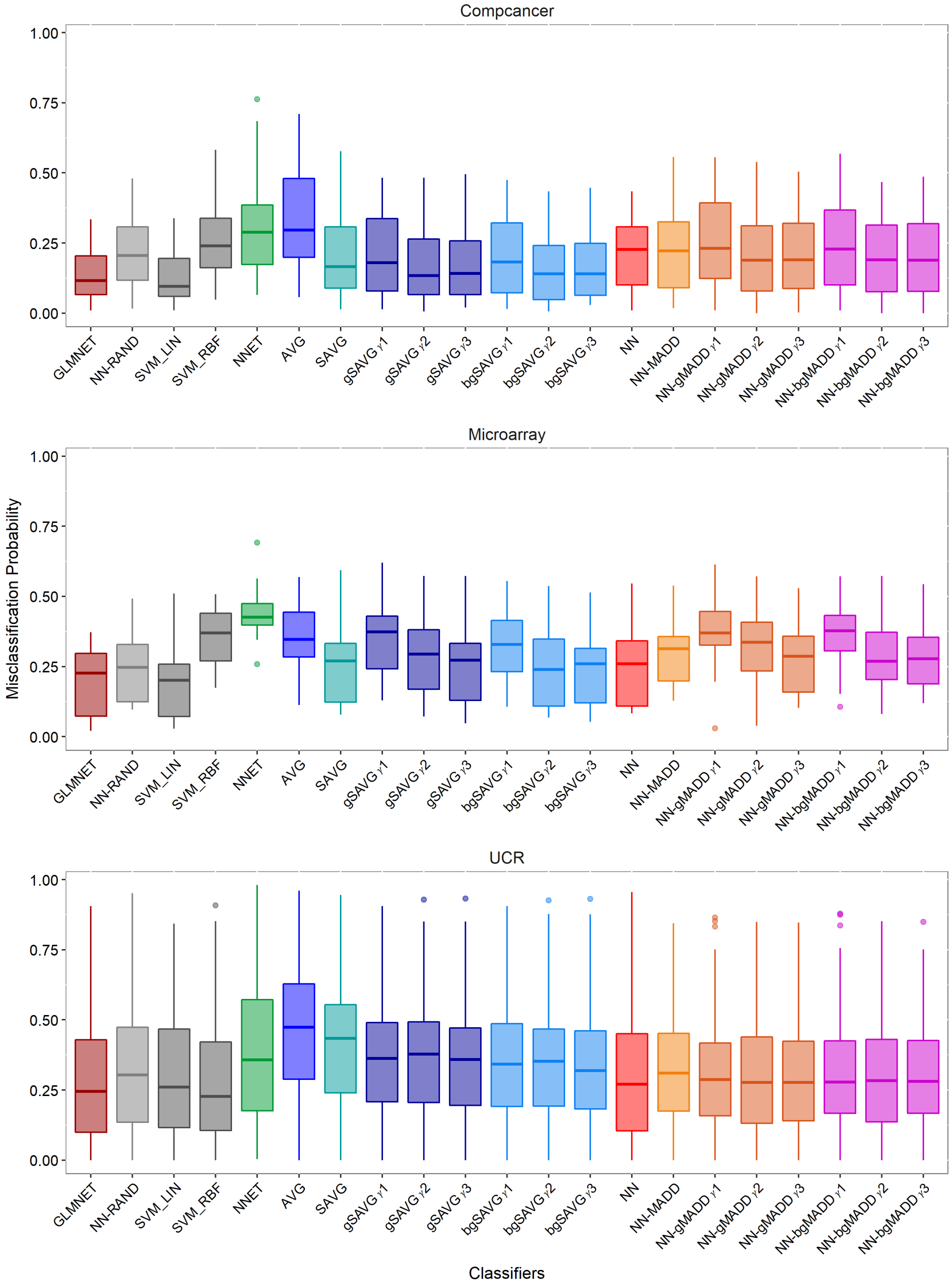

Figure 10: Boxplot of the estimated misclassification rates corresponding to various classifiers in the Compcancer, Microarray and UCR databases.

The UCR data archive is quite diverse with $d \sim 24 - 2700$ and $n \sim 20 - 700$. The number of classes J varies from 2 to 52. Again, GLMNET invokes dimension reduction by identifying sparse components, and yields the best performance. Performance of SVM-RBF improves substantially in this database. The NN-bgMADD classifier also performs quite well and secures a competitive position. Linear classifiers like GLMNET and SVM-LIN perform quite well in data sets with clear differences in their locations, while popular non-linear classifiers like SVM-RBF and NN yield good performance in data sets with difference in scales and/or shapes. In particular, GLMNET and SVM-LIN outperform the non-linear classifiers in the `Coffee` and `Wine` data sets, while SVM-RBF and NN outperform the linear classifiers in the `CinCECGtorso`, `MoteStrain` and `Synthetic Control` data sets. The NN-bgMADD classifiers seem to have a slight edge over the corresponding bgSAVG classifiers here. Generally, we observe a large variability in the boxplots for the UCR database because of the presence of data sets with very high as well as low misclassification rates. In particular, the `PigAirwayPressure` data with 52 classes has a misclassification rate of more than 80% across all classifiers, whereas we obtain *perfect classification* for these classifiers in the `InsectEPGRgularTrain` data with 3 classes.

7. Concluding Remarks

In this article, we have studied the HDLSS asymptotic properties of some distance based classifiers. We have analyzed and generalized the popular average distance classifier and the nearest neighbor classifier. On a theoretical note, we have proved that the misclassification probability of the generalized classifiers go to zero (i.e., *perfect classification*) in the HDLSS asymptotic regime under very general conditions. Using a variety of simulated examples and real data sets from three databases, we have amply demonstrated improved performance of the proposed classifiers when compared with a wide variety of popular classifiers.

The idea of clustering of components in Section 3 allows us to theoretically explore several possible ways in which d can grow to infinity. In this work, we have considered the case where the block sizes are bounded, while the number of blocks increases with the dimension. One can also keep the number of blocks fixed and allow the size of some (or, all) blocks to grow with d. This may lead to concentration of distances within blocks, and the proposed classifiers will then face issues similar to those discussed in Hall et al. (2005). The remaining possibility is to allow both the number of blocks as well as sizes of the blocks to grow to infinity. This, of course, is a complicated setup for theoretical analysis and out of the scope of this article.

Another aspect is handling sparsity in the feature variables. In our theoretical investigations for the generalized classifiers, assumption $(A3)$ corresponds to the case when the number of informative components scales as d, but this can be relaxed further (see Sarkar et al. (2020) for more details). In particular, if the variables are weakly dependent, Theorem 2.4 can be proved when the number of informative variables scales as d^{α}, for some $\alpha > 1/2$. A similar remark holds for assumption $(A7)$ in the context of block-generalized classifiers. In practice, however, one would be interested in capturing the sparse structure in a data dependent way and modify the classifiers accordingly. This is a topic of future research.

Acknowledgments

The first and third authors have been partially supported by the DST-SERB grant ECR/2017/000374. The authors would like to thank the Action Editor for his encouragment, and the three anonymous reviewers for their constructive comments and suggestions that substantially improved the paper.

Appendix A. Proofs and Mathematical Details

We begin with proofs of the results stated in Section 3. Proofs of the results in Section 2 are similar, and are in fact special cases (follows by taking $b = d$, equivalently, $d_i = 1$ for $1 \leq i \leq d$) of these proofs. Hence, we omit them.

Proof of Lemma 3.1 Fix $\epsilon > 0$. Let us define $W_i = \gamma({d_i}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2)$ for $1 \leq i \leq b$, where $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$, $1 \leq j, j' \leq J$. Using Chebyshev's inequality, we observe that

$$\mathrm{P}\left[\left|\frac{1}{b}\sum_{i=1}^{b} W_i - \frac{1}{b}\sum_{i=1}^{b}\mathrm{E}(W_i)\right| > \epsilon\right] \leq \frac{1}{\epsilon^2}\mathrm{E}\left[\frac{1}{b}\sum_{i=1}^{b} W_i - \frac{1}{b}\sum_{i=1}^{b}\mathrm{E}(W_i)\right]^2.$$

We are going to show

$$\mathrm{E}\left[\frac{1}{b}\sum_{i=1}^{b} W_i - \frac{1}{b}\sum_{i=1}^{b}\mathrm{E}(W_i)\right]^2 = \mathrm{Var}\left[\frac{1}{b}\sum_{i=1}^{b} W_i\right] \to 0 \text{ as } b \to \infty.$$

Observe that

$$\begin{aligned} 0 \leq\ & \mathrm{Var}\big[b^{-1}\sum_{i=1}^{b} W_i\big] && \text{(A.1)}\\ &= b^{-2}\sum_{i=1}^{b}\mathrm{Var}\big[W_i\big] + 2b^{-2}\sum\sum_{1\leq i<i'\leq b}\mathrm{Cov}\left(W_i, W_{i'}\right)\\ &= b^{-2}\sum_{i=1}^{b}\mathrm{Var}\big[W_i\big] + 2b^{-2}\sum\sum_{1\leq i<i'\leq b}\mathrm{Corr}\left(W_i, W_{i'}\right)\sqrt{\mathrm{Var}[W_i]\mathrm{Var}[W_{i'}]}\\ &\leq b^{-2}\sum_{i=1}^{b}\mathrm{E}[W_i^2] + 2b^{-2}\sum\sum_{1\leq i<i'\leq b}\mathrm{Corr}\left(W_i, W_{i'}\right)\sqrt{\mathrm{E}[W_i^2]\mathrm{E}[W_{i'}^2]}\\ &\leq b^{-2}\sum_{i=1}^{b} c_2 + 2c_2 b^{-2}\sum\sum_{1\leq i<i'\leq b}\mathrm{Corr}\left(W_i, W_{i'}\right) \quad \text{[by assumption } (A5)]\\ &\leq c_2 b^{-1} + 2c_2 b^{-2}\sum\sum_{1\leq i<i'\leq b}\mathrm{Corr}\left(W_i, W_{i'}\right).\\ &=\ o(1) \text{ [by assumption } (A6)]. && \text{(A.2)} \end{aligned}$$

Therefore, $\left|b^{-1}\sum_{i=1}^{b} W_i - b^{-1}\sum_{i=1}^{b}\mathrm{E}[W_i]\right| \xrightarrow{P} 0$ as $b \to \infty$. Since ϕ is uniformly continuous, it follows from the definition of uniform continuity that for any $\epsilon_1 > 0$, there exists $\epsilon_2 > 0$ such that

$$\mathrm{P}\big[\big|b^{-1}\sum_{i=1}^{b} W_i - b^{-1}\sum_{i=1}^{b}\mathrm{E}[W_i]\big| \leq \epsilon_2\big] \leq \mathrm{P}\big[\big|\phi(b^{-1}\sum_{i=1}^{b} W_i) - \phi(b^{-1}\sum_{i=1}^{b}\mathrm{E}[W_i])\big| \leq \epsilon_1\big].$$

Since, $\lim_{b\to\infty}\mathrm{P}\big[\big|b^{-1}\sum_{i=1}^{b} W_i - b^{-1}\sum_{i=1}^{b}\mathrm{E}[W_i]\big| \leq \epsilon_2\big] = 1$,

$$\left|\phi(b^{-1}\sum_{i=1}^{b} W_i) - \phi(b^{-1}\sum_{i=1}^{b}\mathrm{E}[W_i])\right| \xrightarrow{P} 0 \text{ as } b \to \infty.$$

Hence, $\big|h_b(\mathbf{U},\mathbf{V}) - \tilde{h}_b(j,j')\big| \xrightarrow{P} 0$ as $b \to \infty$ for all $1 \le j, j' \le J$. ■

Proof of Corollary 3.2 It follows from Lemma 3.1 that for independent random vectors $\mathbf{Z} \sim \mathbf{F}_j$, and $\mathbf{X}, \mathbf{X}' \overset{i.i.d.}{\sim} \mathbf{F}_{j'}$ with $1 \le j, j' \le J$, we have

$$|h_b(\mathbf{Z},\mathbf{X}) - \tilde{h}_b(j,j')| \xrightarrow{P} 0 \text{ and } |h_b(\mathbf{X},\mathbf{X}') - \tilde{h}_b(j',j')| \xrightarrow{P} 0 \text{ as } b \to \infty.$$

This further implies that

$$\begin{aligned}&\Bigg|n_{j'}^{-1} \sum_{\mathbf{X} \in \mathscr{X}_j} h_b(\mathbf{Z},\mathbf{X}) - \tilde{h}_b(j,j')\Bigg| \xrightarrow{P} 0 \text{ and}\\ &\Bigg|\{n_{j'}(n_{j'}-1)\}^{-1} \sum_{\mathbf{X},\mathbf{X}' \in \mathscr{X}_j} h_b(\mathbf{X},\mathbf{X}') - \tilde{h}_b(j',j')\Bigg| \xrightarrow{P} 0 \text{ as } b \to \infty. \qquad \text{(A.3)}\end{aligned}$$

(a) Recall that for any $1 \le j, j' \le J$,

$$\begin{aligned}\xi_{jb}(\mathbf{Z}) &= n_j^{-1} \sum_{\mathbf{X} \in \mathscr{X}_j} h_b(\mathbf{Z},\mathbf{X}) - \{2n_j(n_j-1)\}^{-1} \sum_{\mathbf{X},\mathbf{X}' \in \mathscr{X}_j} h_b(\mathbf{X},\mathbf{X}'),\\ \xi_{j'b}(\mathbf{Z}) &= n_{j'}^{-1} \sum_{\mathbf{X} \in \mathscr{X}_{j'}} h_b(\mathbf{Z},\mathbf{X}) - \{2n_{j'}(n_{j'}-1)\}^{-1} \sum_{\mathbf{X},\mathbf{X}' \in \mathscr{X}_{j'}} h_b(\mathbf{X},\mathbf{X}'), \text{ and}\\ \tilde{\xi}_b(j,j') &= \tilde{h}_b(j,j') - \frac{1}{2}\big(\tilde{h}_b(j',j') + \tilde{h}_b(j,j)\big).\end{aligned}$$

Since $\mathbf{Z} \sim \mathbf{F}_j$, it follows from (A.3) that

$$|\xi_{j'b}(\mathbf{Z}) - \{\tilde{h}_b(j,j') - \tilde{h}_b(j',j')/2\}\big| \xrightarrow{P} 0 \text{ and } |\xi_{jb}(\mathbf{Z}) - \tilde{h}_b(j,j)/2\big| \xrightarrow{P} 0 \text{ as } b \to \infty. \quad \text{(A.4)}$$

Consequently,

$$\begin{aligned}&\Bigg|\big\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\big\} - \Big\{\tilde{h}_b(j,j') - \frac{1}{2}\big(\tilde{h}_b(j',j') + \tilde{h}_b(j,j)\big)\Big\}\Bigg| \xrightarrow{P} 0 \text{ as } b \to \infty\\ \implies &\big|\big\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\big\} - \tilde{\xi}_b(j,j')\big| \xrightarrow{P} 0 \text{ as } b \to \infty.\end{aligned}$$

(b) Recall that $\mathbf{Z} \sim \mathbf{F}_j$ and $\mathbf{X} \sim \mathbf{F}_{j'}$ with $1 \le j, j' \le J$, and $\psi_b(\mathbf{Z},\mathbf{X})$ can be expressed as follows:

$$\frac{1}{n-1}\Bigg(\sum_{\mathbf{X}' \in \mathscr{X}_{j'} \setminus \{\mathbf{X}\}} \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| + \sum_{\mathbf{X}' \in \mathscr{X} \setminus \mathscr{X}_{j'}} \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big|\Bigg).$$

Now, using triangle inequality (repeatedly), we obtain

$$0 \le \big|\psi_b(\mathbf{Z},\mathbf{X}) - \tilde{\tau}_b(j,j')\big|$$

$$
\begin{aligned}
&= \Bigg| \frac{1}{n-1}\Bigg\{ \sum_{\mathbf{X}'\in\mathscr{X}_{j'}\setminus\{\mathbf{X}\}} \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| + \sum_{\mathbf{X}'\in\mathscr{X}\setminus\mathscr{X}_{j'}} \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| \Bigg\} \\
&\quad - \Bigg\{ \frac{n_{j'}-1}{n-1} \mid \tilde{h}_b(j,j') - \tilde{h}_b(j',j') \mid + \sum_{l\neq j'} \frac{n_l}{n-1} \mid \tilde{h}_b(j,l) - \tilde{h}_b(j',l) \mid \Bigg\} \Bigg| \\
&= \Bigg| \frac{1}{n-1}\Bigg\{ \sum_{\mathbf{X}'\in\mathscr{X}_{j'}\setminus\{\mathbf{X}\}} \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| - (n_{j'}-1) \mid \tilde{h}_b(j,j') - \tilde{h}_b(j',j') \mid \\
&\quad + \sum_{\mathbf{X}'\in\mathscr{X}\setminus\mathscr{X}_{j'}} \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| - \sum_{l\neq j'} \frac{n_l}{n-1} \mid \tilde{h}_b(j,l) - \tilde{h}_b(j',l) \mid \Bigg\} \Bigg| \\
&\leq \frac{1}{n-1}\Bigg\{ \sum_{\mathbf{X}'\in\mathscr{X}_{j'}\setminus\{\mathbf{X}\}} \Big| \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| - \mid \tilde{h}_b(j,j') - \tilde{h}_b(j',j') \mid \Big| \\
&\quad + \sum_{l\neq j'} \sum_{\mathbf{X}'\in\mathscr{X}_l} \Big| \big|h_b(\mathbf{Z},\mathbf{X}') - h_b(\mathbf{X},\mathbf{X}')\big| - \mid \tilde{h}_b(j,l) - \tilde{h}_b(j',l) \mid \Big| \Bigg\} \\
&\leq \frac{1}{n-1}\Bigg\{ \sum_{\mathbf{X}'\in\mathscr{X}_{j'}\setminus\{\mathbf{X}\}} \big|h_b(\mathbf{Z},\mathbf{X}') - \tilde{h}_b(j,j')\big| + \sum_{\mathbf{X}'\in\mathscr{X}_{j'}\setminus\{\mathbf{X}\}} \big|h_b(\mathbf{X},\mathbf{X}') - \tilde{h}_b(j',j')\big| \\
&\quad + \sum_{l\neq j'} \sum_{\mathbf{X}'\in\mathscr{X}_l} \big|h_b(\mathbf{Z},\mathbf{X}') - \tilde{h}_b(j,l)\big| + \sum_{l\neq j'} \sum_{\mathbf{X}'\in\mathscr{X}_l} \big|h_b(\mathbf{X},\mathbf{X}') - \tilde{h}_b(j',l)\big| \Bigg\}.
\end{aligned}
$$

It follows from Lemma 3.1 that each of the summands converge to 0 in probability as $b\to\infty$. Therefore, for a fixed sample size n, $\big|\psi_b(\mathbf{Z},\mathbf{X}) - \tilde{\tau}_b(j,j')\big| \xrightarrow{P} 0$ as $b\to\infty$ for all $1\leq j,j'\leq J$.

Let us assume that $j\neq j'$. We have $\tau_{jb}(\mathbf{Z}) = \min_{\mathbf{X}\in\mathscr{X}_j}\psi_b(\mathbf{Z},\mathbf{X})$, and $\tau_{j'b}(\mathbf{Z}) = \min_{\mathbf{X}\in\mathscr{X}_{j'}}\psi_b(\mathbf{Z},\mathbf{X})$. Since $\mathbf{Z}\sim\mathbf{F}_j$, we get

$$\big|\tau_{j'b}(\mathbf{Z}) - \tilde{\tau}_b(j,j')\big| \xrightarrow{P} 0 \text{ and} \big|\tau_{jb}(\mathbf{Z}) - \tilde{\tau}_b(j,j)\big| \xrightarrow{P} 0 \text{ as } b\to\infty.$$

Since $\tilde{\tau}_b(j,j)=0$, it follows that

$$\big|\{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} - \tilde{\tau}_b(j,j')\big| \xrightarrow{P} 0 \text{ as } b\to\infty.$$

■

Proof of Lemma 3.3 Suppose that $\mathbf{X}_1,\mathbf{X}_2$ are i.i.d. copies of $\mathbf{X}\sim\mathbf{F}_j$, and $\mathbf{X}_3,\mathbf{X}_4$ are i.i.d. copies of $\mathbf{X}'\sim\mathbf{F}_{j'}$ for $1\leq j\neq j'\leq J$. Let us denote $\tilde{h}_b(j,j)=\phi(A_{1b})$, $\tilde{h}_b(j',j')=\phi(A_{2b})$ and $\tilde{h}_b(j,j')=\phi(A_{3b})$, where $A_{1b}=b^{-1}\sum_{i=1}^{b}\mathrm{E}\big[\gamma(d_i^{-1}\|\mathbf{X}_{1i}-\mathbf{X}_{2i}\|^2)\big]$, $A_{2b}=b^{-1}\sum_{i=1}^{b}\mathrm{E}\big[\gamma(d_i^{-1}\|\mathbf{X}_{3i}-\mathbf{X}_{4i}\|^2)\big]$ and $A_{3b}=b^{-1}\sum_{i=1}^{b}\mathrm{E}\big[\gamma(d_i^{-1}\|\mathbf{X}_{1i}-\mathbf{X}_{3i}\|^2)\big]$.

(a) For $1\leq i\leq b$ and $1\leq j\neq j'\leq J$, we have

$$e(\mathbf{F}_{j,i},\mathbf{F}_{j',i}) = \mathrm{E}[\gamma(d_i^{-1}\|\mathbf{X}_{1i}-\mathbf{X}_{3i}\|^2)] - \frac{1}{2}\Big\{\mathrm{E}[\gamma(d_i^{-1}\|\mathbf{X}_{1i}-\mathbf{X}_{2i}\|^2)] + \mathrm{E}[\gamma(d_i^{-1}\|\mathbf{X}_{3i}-\mathbf{X}_{4i}\|^2)]\Big\}$$

is the energy distance between the distributions $\mathbf{F}_{j,i}$ and $\mathbf{F}_{j',i}$. Baringhaus and Franz (2010) showed that the energy distance between two distributions is always non-negative, i.e., $e(\mathbf{F}_{j,i}, \mathbf{F}_{j',i}) \geq 0$, for all $1 \leq i \leq b$ and $1 \leq j \neq j' \leq J$. Therefore,

$$\mathrm{E}[\gamma(d_i^{-1}\|\mathbf{X}_{1i}-\mathbf{X}_{3i}\|^2)] \geq \frac{1}{2}\Big\{\mathrm{E}[\gamma(d_i^{-1}\|\mathbf{X}_{1i}-\mathbf{X}_{2i}\|^2)]+\mathrm{E}[\gamma(d_i^{-1}\|\mathbf{X}_{3i}-\mathbf{X}_{4i}\|^2)]\Big\},\ \forall 1 \leq i \leq b.$$

This implies that $A_{3b} \geq \frac{1}{2}(A_{1b} + A_{2b})$. Since ϕ is increasing and concave, we have $\phi(A_{3b}) \geq \phi\big(\frac{1}{2}A_{1b} + \frac{1}{2}A_{2b}\big) \geq \frac{1}{2}\phi(A_{1b}) + \frac{1}{2}\phi(A_{2b})$. This further implies that $\tilde{\xi}_b(j,j') = \tilde{h}_b(j,j') - \frac{1}{2}\big\{\tilde{h}_b(j,j) + \tilde{h}_b(j',j')\big\} \geq 0$.

Baringhaus and Franz (2010) also showed that $e(\mathbf{F}_{j,i}, \mathbf{F}_{j',i}) = 0$ if and only if $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$, and we have $\tilde{\xi}_b(j,j') = 0$. So, we have $\phi(A_{3b}) = \frac{1}{2}\phi(A_{1b}) + \frac{1}{2}\phi(A_{2b})$. Since ϕ is concave and increasing, it is straightforward to check that $\frac{1}{2}A_{1b} + \frac{1}{2}A_{2b} \geq A_{3b}$. But, we already know that $A_{3b} \geq \frac{1}{2}A_{1b} + \frac{1}{2}A_{2b}$ and hence, the equality follows.

This further implies that $\frac{1}{b}\sum_{i=1}^{b} e(\mathbf{F}_{j,i}, \mathbf{F}_{j',i}) = 0$ for all $1 \leq j \neq j' \leq J$, i.e., $e(\mathbf{F}_{j,i}, \mathbf{F}_{j',i}) = 0$ for all $1 \leq i \leq b$ and $1 \leq j \neq j' \leq J$. Clearly, $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$ for all $1 \leq i \leq b$ and $1 \leq j \neq j' \leq J$ now follows.

Let us assume that $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$ for all $1 \leq i \leq b$ and $1 \leq j \neq j' \leq J$. Therefore, we get

$$\mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{1i} - \mathbf{X}_{2i}\|^2\big)\big] = \mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{1i} - \mathbf{X}_{3i}\|^2\big)\big] = \mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{3i} - \mathbf{X}_{4i}\|^2\big)\big]$$

which implies that $A_{1b} = A_{2b} = A_{3b}$. As a consequence, we obtain $\tilde{h}_b(j,j) = \tilde{h}_b(j,j') = \tilde{h}_b(j',j')$, and hence $\tilde{\xi}_b(j,j') = 0$ for $1 \leq j \neq j' \leq J$.

(b) Recall that for $1 \leq j \neq j' \leq J$, we have

$$\tilde{\tau}_b(j,j') = \frac{n_{j'}-1}{n-1} \mid \tilde{h}_b(j,j') - \tilde{h}_b(j',j') \mid + \sum_{l \neq j'} \frac{n_l}{n-1} \mid \tilde{h}_b(j,l) - \tilde{h}_b(j',l) \mid \geq 0.$$

If $\tilde{\tau}_b(j,j') = 0$, then $\tilde{h}_b(j,l) = \tilde{h}_b(j',l)$ for all $1 \leq l \leq J$. So, we get $\tilde{h}_b(j,j) = \tilde{h}_b(j,j') = \tilde{h}_b(j',j')$ $[\because \tilde{h}_b(j,j') = \tilde{h}_b(j',j)]$. This further implies $\phi(A_{1b}) = \phi(A_{2b}) = \phi(A_{3b})$, and since ϕ is one-to-one, we get $A_{1b} = A_{2b} = A_{3b}$. So, we have $\tilde{\xi}_b(j,j') = A_{3b} - \frac{1}{2}\{A_{1b} + A_{2b}\} = 0$. This implies $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$ for all $1 \leq i \leq b$.

Let us now assume that $\mathbf{F}_{j,i} = \mathbf{F}_{j',i}$ for all $1 \leq i \leq b$. Consequently, for $\mathbf{X}' \sim \mathbf{F}_l$ with $1 \leq l \leq J$ and $1 \leq i \leq b$, we get the following

$$\begin{aligned}
&\mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{1i} - \mathbf{X}'_i\|^2\big)\big] = \mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{3i} - \mathbf{X}'_i\|^2\big)\big]\\
\implies &\phi\bigg(b^{-1}\sum_{i=1}^{b}\mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{1i} - \mathbf{X}'_i\|^2\big)\big]\bigg) = \phi\bigg(b^{-1}\sum_{i=1}^{b}\mathrm{E}\big[\gamma\big({d_i}^{-1}\|\mathbf{X}_{3i} - \mathbf{X}'_i\|^2\big)\big]\bigg)\\
\implies &\tilde{\tau}_b(j,j') = 0.
\end{aligned}$$

This completes the proof. ■

Recall that assumption $(A7)$ implies $\liminf_{b\to\infty} \tilde{\tau}_b(j,j') > 0$ for any $1 \leq j \neq j' \leq J$. We now state and prove this fact below.

Lemma 1 *If* $\liminf_{b\to\infty} \tilde{\xi}_b^{\phi,\gamma}(j,j') > 0$, *then we have* $\liminf_{b\to\infty} \tilde{\tau}_b^{\phi,\gamma}(j,j') > 0$ *for any* $1 \leq j \neq j' \leq J$.

Proof of Lemma 1 Recall that

$$\tilde{\xi}_b(j,j') = \tilde{h}_b(j,j') - \frac{1}{2}\big[\tilde{h}_b(j,j) + \tilde{h}_b(j',j')\big], \text{ and}$$
$$\tilde{\tau}_b(j,j') = \sum_{l\neq j'}\left\{\frac{n_l}{n-1}|\tilde{h}_b(j,l) - \tilde{h}_b(j',l)|\right\} + \frac{n_{j'}-1}{n-1}|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')|.$$

Since

$$\begin{aligned}\tilde{\xi}_b(j,j') &= \tilde{h}_b(j,j') - \frac{1}{2}\big[\tilde{h}_b(j,j) + \tilde{h}_b(j',j')\big]\\ &= \frac{1}{2}\big[\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big] + \frac{1}{2}\big[\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big]\\ &\leq \frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big| + \frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big|,\end{aligned}$$

it follows that

$$\liminf_{b\to\infty}\tilde{\xi}_b(j,j') > 0 \implies \liminf_{b\to\infty}\left(\frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big| + \frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big|\right) > 0.$$

Now, let us assume that

$$\liminf_{b\to\infty}\left(\frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big| + \frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big|\right) = c,$$

for some $c > 0$. This means that for any $\epsilon > 0$, there exists a $b(\epsilon)$ such that for all $b \geq b(\epsilon)$, we have

$$\begin{aligned}&\frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big| + \frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big| > c - \epsilon\\ \implies &\frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big| > \frac{c-\epsilon}{2}, \text{ or } \frac{1}{2}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big| > \frac{c-\epsilon}{2}\\ \implies &\frac{n_j}{n-1}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j,j)\big| + \frac{n_{j'}-1}{n-1}\big|\tilde{h}_b(j,j') - \tilde{h}_b(j',j')\big|\\ &> \min\left\{\frac{n_j(c-\epsilon)}{n-1}, \frac{(n_{j'}-1)(c-\epsilon)}{n-1}\right\}\\ \implies &\tilde{\tau}_b(j,j') > \min\left\{\frac{n_j(c-\epsilon)}{n-1}, \frac{(n_{j'}-1)(c-\epsilon)}{n-1}\right\}.\end{aligned}$$

Since ϵ is chosen arbitrarily, we obtain the following

$$\liminf_{b\to\infty}\tilde{\tau}_b(j,j') > c\min\left\{\frac{n_j}{n-1}, \frac{n_{j'}-1}{n-1}\right\} > 0.$$

Similarly, it can be shown that

$$\liminf_{b\to\infty}\tilde{\tau}_b(j',j) > c\min\left\{\frac{n_j-1}{n-1}, \frac{n_{j'}}{n-1}\right\} > 0.$$

This completes the proof. ■

Proof of Theorem 3.4

(a) The misclassification probability of the bgSAVG classifier is defined as

$$\Delta_{\text{bgSAVG}} = \mathrm{P}[\delta_{\text{bgSAVG}}(\mathbf{Z}) \neq Y],$$

where Y denotes the true label of $\mathbf{Z}$. We will prove that $\Delta_{\text{bgSAVG}} \to 0$ as $b \to \infty$. Now, note that

$$\begin{aligned}
0 &\leq \lim_{b\to\infty} \mathrm{P}[\delta_{\text{bgSAVG}}(\mathbf{Z}) \neq Y] \\
&= \lim_{b\to\infty} \sum_{j=1}^{J} \mathrm{P}[\delta_{\text{bgSAVG}}(\mathbf{Z}) \neq j, \mathbf{Z} \sim \mathbf{F}_j] \\
&= \sum_{j=1}^{J} \pi_j \lim_{b\to\infty} \mathrm{P}[\delta_{\text{bgSAVG}}(\mathbf{Z}) \neq j \mid \mathbf{Z} \sim \mathbf{F}_j] \\
&= \sum_{j=1}^{J} \pi_j \lim_{b\to\infty} \mathrm{P}[\xi_{jb}(\mathbf{Z}) - \xi_{j'b}(\mathbf{Z}) > 0 \text{ for some } j' \neq j, 1 \leq j' \leq J \mid \mathbf{Z} \sim \mathbf{F}_j] \\
&\leq \sum_{j=1}^{J} \pi_j \lim_{b\to\infty} \sum_{1\leq j\neq j'\leq J} \mathrm{P}[\xi_{jb}(\mathbf{Z}) - \xi_{j'b}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_j] \\
&= \sum_{j=1}^{J} \pi_j \sum_{1\leq j\neq j'\leq J} \lim_{b\to\infty} \mathrm{P}[\xi_{jb}(\mathbf{Z}) - \xi_{j'b}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_j]. \qquad \text{(A.5)}
\end{aligned}$$

For any $\theta > 0$ and $\epsilon > 0$, there exists a B_1 such that for all $b \geq B_1$, we have

$$\begin{aligned}
&\mathrm{P}[|\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) - \tilde{\xi}_b(j,j')| < \theta \mid \mathbf{Z} \sim \mathbf{F}_j] > 1-\epsilon \text{ [see Corollary 3.2(a)]} \\
\implies &\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) - \tilde{\xi}_b(j,j') > -\theta \mid \mathbf{Z} \sim \mathbf{F}_j] > 1-\epsilon \\
\implies &\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > -\theta + \tilde{\xi}_b(j,j') \mid \mathbf{Z} \sim \mathbf{F}_j] > 1-\epsilon.
\end{aligned}$$

Let $\liminf_b \tilde{\xi}_b(j,j')$ be denoted by $\tilde{\xi}(j,j')$. For any $\theta' > 0$, there exists a B' such that $\tilde{\xi}_b(j,j') > \xi(j,j') - \theta'$ for all $b \geq B'$. Therefore,

$$\begin{aligned}
&\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > -\theta + \tilde{\xi}_b(j,j') \mid \mathbf{Z} \sim \mathbf{F}_j] \\
&\quad \leq \mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > -\theta - \theta' + \tilde{\xi}(j,j') \mid \mathbf{Z} \sim \mathbf{F}_j] \text{ for all } b \geq B' \\
\implies &\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > -\theta - \theta' + \tilde{\xi}(j,j') \mid \mathbf{Z} \sim \mathbf{F}_j] > 1-\epsilon \text{ for all } b \geq \max\{B', B_1\}. \qquad \text{(A.6)}
\end{aligned}$$

Since θ, θ' are arbitrary, it can be concluded from equation (A.6) that

$$\begin{aligned}
&\lim_{b\to\infty} \mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) \geq \tilde{\xi}(j,j') \mid \mathbf{Z} \sim \mathbf{F}_j] = 1 \\
\implies &\lim_{b\to\infty} \mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_j] = 1 \; [\because \tilde{\xi}(j,j') > 0]
\end{aligned}$$

$$\implies \lim_{b\to\infty} \mathrm{P}[\xi_{jb}(\mathbf{Z}) - \xi_{j'b}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_j] = 0. \tag{A.7}$$

Now, it follows from equations (A.5) and (A.7) that

$$\lim_{b\to\infty} \mathrm{P}[\delta_{\mathrm{bgSAVG}}(\mathbf{Z}) \neq Y] = \sum_{j=1}^{J} \pi_j \cdot 0 = 0.$$

(b) Proof for the misclassification probability of the NN-bgMADD classifier is similar, and follows along the lines of the proof of part (a). Please check Section 1 of the Supplementary for a proof. ■

Proof of Theorem 3.5 Suppose $0 \leq s_i, t_i \leq 1$, for $1 \leq i \leq K$. Then

$$\begin{aligned}
\prod_{i=1}^{K}(s_i + t_i) &= \sum_{S\subseteq\{1,\dots,K\}} \prod_{i\in S} s_i \prod_{i\in\{1,\dots,K\}\setminus S} t_i \\
&= \prod_{i=1}^{K} s_i + \sum_{S\subset\{1,\dots,K\}} \prod_{i\in S} s_i \prod_{i\in\{1,\dots,K\}\setminus S} t_i \\
&\leq \prod_{i=1}^{K} s_i + \sum_{S\subset\{1,\dots,K\}} \prod_{i\in\{1,\dots,K\}\setminus S} t_i \\
&\leq \prod_{i=1}^{K} s_i + \sum_{i\in\{1,\dots,K\}} C_K\ t_i, \qquad (A.8)
\end{aligned}$$

for some appropriate constant $C_K > 0$.
Recall that $\Delta_{\mathrm{bgSAVG}} = 1 - \mathrm{P}[\delta_{\mathrm{bgSAVG}}(\mathbf{Z}) = Y]$ and $\Delta_{\mathrm{NN-bgMADD}} = 1 - \mathrm{P}[\delta_{\mathrm{NN-bgMADD}}(\mathbf{Z}) = Y]$. Here,

$$\mathrm{P}[\delta_{\mathrm{bgSAVG}}(\mathbf{Z}) = Y] = \sum_{j=1}^{J} \pi_j \mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0\ \forall j' \neq j, 1 \leq j' \leq J|\mathbf{Z} \sim \mathbf{F}_j], \text{ and}$$

$$\mathrm{P}[\delta_{\mathrm{NN-bgMADD}}(\mathbf{Z}) = Y] = \sum_{j=1}^{J} \pi_j \mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0\ \forall j' \neq j, 1 \leq j' \leq J|\mathbf{Z} \sim \mathbf{F}_j].$$

It is to be noted that given $\mathbf{Z}$ and $\mathscr{X}_j$ (training data of the j-th class), $\tau_{kb}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})$ and $\tau_{lb}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})$ are independently distributed for all $1 \leq k \neq l \leq J, k, l \neq j$. Therefore, for any $1 \leq j \leq J$, we can write the following

$$\begin{aligned}
&\mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0\ \forall j' \neq j, 1 \leq j' \leq J|\mathbf{Z} \sim \mathbf{F}_j] \\
&= \mathrm{E}\big\{\mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0\ \forall j' \neq j, 1 \leq j' \leq J|\mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\big\} \\
&= \mathrm{E}\Bigg\{\prod_{1\leq j'\neq j\leq J} \mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0|\mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\Big\{ \prod_{1\leq j'\neq j\leq J} \big(\mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0,\ \xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j] \\
&\qquad + \mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0, \xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) < 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\big)\Big\} \\
&\leq \mathrm{E}\Big\{ \prod_{1\leq j'\neq j\leq J} \big(\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j] \\
&\qquad + \mathrm{P}[\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} < 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\big)\Big\} \\
&\leq \mathrm{E}\Big\{ \prod_{1\leq j'\neq j\leq J} \big(\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\big) \\
&\qquad + \sum_{1\leq j'\neq j\leq J} C_J \cdot \mathrm{P}[\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} < 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\Big\} \text{ [using (A.8)]} \\
&= \mathrm{E}\Big\{ \prod_{1\leq j'\neq j\leq J} \big(\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\big)\Big\} \\
&\qquad + \ \mathrm{E}\Big\{ \sum_{1\leq j'\neq j\leq J} C_J \cdot \mathrm{P}[\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} < 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\Big\} \\
&= \mathrm{E}\Big\{ \prod_{1\leq j'\neq j\leq J} \big(\mathrm{P}[\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z}) > 0 | \mathbf{Z} \sim \mathbf{F}_j, \mathscr{X}_j]\big)\Big\} \\
&\qquad + \sum_{1\leq j'\neq j\leq J} C_J \cdot \mathrm{P}[\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} < 0 | \mathbf{Z} \sim \mathbf{F}_j]. \qquad \text{(A.9)}
\end{aligned}
$$

For $\mathbf{Z} \sim \mathbf{F}_j$ and $1 \leq j' \neq j \leq J$, using Corollary 3.2, we have

$$\big|\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \tilde{\xi}_b(j, j')\big| \xrightarrow{\mathrm{P}} 0 \text{ and } \big|\{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} - \tilde{\tau}_b(j, j')\big| \xrightarrow{\mathrm{P}} 0 \text{ as } b \to \infty.$$

This now implies that

$$\big|\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} - \{\tilde{\xi}_b(j, j') - \tilde{\tau}_b(j, j')\}\big| \xrightarrow{\mathrm{P}} 0 \text{ as } b \to \infty.$$

Therefore, for any $\theta > 0,\ \epsilon > 0$ and j there exists a $B_{j,j'}$ such that for all $b \geq B_{j,j'}$

$$\mathrm{P}[\big|\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} - \{\tilde{\xi}_b(j, j') - \tilde{\tau}_b(j, j')\}\big| < \theta \big| \mathbf{Z} \sim F_j] > 1 - \epsilon.$$

We assume $\tilde{\xi}_b(j, j') > \tilde{\tau}_b(j, j')$ for all $b \geq B_1$ and $1 \leq j \neq j' \leq J$. Let $\theta_0 = \liminf_b \big(\tilde{\xi}_b(j, j') - \tilde{\tau}_b(j, j')\big)$. By assumption $(A9)$, $\theta_0 > 0$. Hence, for any $0 < \theta < \theta_0$ and $\epsilon > 0$, there exists a $b'(\theta_0, \theta, \epsilon)$ such that for all $b \geq b'(\theta_0, \theta, \epsilon)$

$$\mathrm{P}[\{\xi_{j'b}(\mathbf{Z}) - \xi_{jb}(\mathbf{Z})\} - \{\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z})\} \leq 0 \big| \mathbf{Z} \sim F_j] < \epsilon.$$

From equation (A.9), we now obtain

$$\mathrm{P}[\tau_{j'b}(\mathbf{Z}) - \tau_{jb}(\mathbf{Z}) > 0\ \forall j \neq j' | \mathbf{Z} \sim \mathbf{F}_j]$$

$$\begin{aligned}
&\leq \mathrm{E}\Big\{\prod_{1\leq j\neq j'\leq J} \mathrm{P}[\xi_{j'b}(\mathbf{Z})-\xi_{jb}(\mathbf{Z})>0|\mathbf{Z}\sim\mathbf{F}_j,\mathscr{X}_j]\Big\}+\sum_{1\leq j\neq j'\leq J} C_J\ \epsilon\\
&=\mathrm{E}\Big\{\prod_{1\leq j\neq j'\leq J} \mathrm{P}[\xi_{j'b}(\mathbf{Z})-\xi_{jb}(\mathbf{Z})>0|\mathbf{Z}\sim\mathbf{F}_j,\mathscr{X}_j]\Big\}+C'_J\ \epsilon\\
&=\mathrm{E}\big\{\mathrm{P}[\xi_{j'b}(\mathbf{Z})-\xi_{jb}(\mathbf{Z})>0\ \forall j'\neq j, 1\leq j'\leq J|\mathbf{Z}\sim\mathbf{F}_j,\mathscr{X}_j]\big\}+C'_J\ \epsilon\\
&=\mathrm{P}[\xi_{j'b}(\mathbf{Z})-\xi_{jb}(\mathbf{Z})>0\ \forall j'\neq j, 1\leq j'\leq J|\mathbf{Z}\sim\mathbf{F}_j]+C'_J\ \epsilon\ \text{ for all } b\geq b'(\theta_0,\theta,\epsilon).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{j=1}^{J}\pi_j\mathrm{P}[\tau_{j'b}(\mathbf{Z})-\tau_{jb}(\mathbf{Z})>0\ \forall j'\neq j, 1\leq j'\leq J|\mathbf{Z}\sim\mathbf{F}_j]\\
&\qquad\leq\sum_{j=1}^{J}\pi_j\mathrm{P}[\xi_{j'b}(\mathbf{Z})-\xi_{jb}(\mathbf{Z})>0\ \forall j'\neq j, 1\leq j'\leq J|\mathbf{Z}\sim\mathbf{F}_j]+C'_J\ \epsilon\\
&\implies \mathrm{P}[\delta_{\mathrm{NN-bgMADD}}(\mathbf{Z})=Y]\leq \mathrm{P}[\delta_{\mathrm{bgSAVG}}(\mathbf{Z})=Y]+C'_J\ \epsilon.
\end{aligned}$$

This now implies that $\Delta_{\mathrm{bgSAVG}}-C'_J\ \epsilon\leq\Delta_{\mathrm{NN-bgMADD}}$ for all $b\geq b'(\theta_0,\theta,\epsilon)$. Since $\epsilon>0$ is arbitrarily, we conclude that

$$\Delta_{\mathrm{bgSAVG}}\leq\Delta_{\mathrm{NN-bgMADD}}\text{ for all } b\geq b'(\theta_0,\theta,\epsilon).$$

Following a similar line of arguments, one can prove that there exist B_1 and B_2 such that if $\tilde{\xi}_b(j,j')<\tilde{\tau}_b(j,j')$ for all $b\geq B_1$ and $1\leq j\neq j'\leq J$, then $\Delta_{\mathrm{bgSAVG}}\geq\Delta_{\mathrm{NN-bgMADD}}$ for all $b\geq B_2$. This completes the proof. ■

Lemma 2 *We now discuss some sufficient conditions for* $\tilde{\xi}^b_{\phi,\gamma}(j,j')\geq(<)\ \tilde{\tau}^b_{\phi,\gamma}(j,j')$ *for* $1\leq j\neq j'\leq J$.
Let us consider a two ($J=2$) class problem. If

i. $\tilde{h}_b(1,2)>\tilde{h}_b(1,1)>\tilde{h}_b(2,2)$ *and* $n_1>n_2+1$,

ii. $\tilde{h}_b(1,2)>\tilde{h}_b(2,2)>\tilde{h}_b(1,1)$ *and* $n_1<n_2-1$,

iii. $\tilde{h}_b(1,1)>\tilde{h}_b(1,2)\geq\frac{3}{4}\tilde{h}_b(1,1)+\frac{1}{4}\tilde{h}_b(2,2)>\tilde{h}_b(2,2)$ *and*

$$n_1>1+\tfrac{n-1}{2}\Big\{\frac{\tilde{h}_b(1,1)-\tilde{h}_b(2,2)}{2\tilde{h}_b(1,2)-\tilde{h}_b(1,1)-\tilde{h}_b(2,2)}\Big\},\ or$$

iv. $\tilde{h}_b(2,2)>\tilde{h}_b(1,2)\geq\frac{1}{4}\tilde{h}_b(1,1)+\frac{3}{4}\tilde{h}_b(2,2)>\tilde{h}_b(1,1)$ *and*

$$n_1<(n-1)\Big\{1-\frac{1}{2}\frac{\tilde{h}_b(2,2)-\tilde{h}_b(1,1)}{2\tilde{h}_b(1,2)-\tilde{h}_b(1,1)-\tilde{h}_b(2,2)}\Big\},$$

then $\tilde{\xi}_b(1,2)>\max\{\tilde{\tau}_b(1,2),\tilde{\tau}_b(2,1)\}$.

Proof of Lemma 2 Please check Section 1 of the Supplementary for a proof. ■

Remark A Assumption ($A8$) holds in various scenarios. In particular, if the component variables of the underlying distributions are i.i.d., then the constants $\tilde{\xi}_b$ and $\tilde{\tau}_b$ are free of b. To realize this, assume $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d}{\sim} \mathbf{F}_1$, $\mathbf{X}_2, \mathbf{X}_4 \overset{i.i.d}{\sim} \mathbf{F}_2$. If $d_i = d_1$, and $\mathbf{X}_{1i} \overset{i.i.d}{\sim} \mathbf{F}_{1,i}$, $\mathbf{X}_{3i} \overset{i.i.d}{\sim} \mathbf{F}_{2,i}$ for all $1 \leq i \leq b$, then we have

$$\begin{aligned}\tilde{h}_b(1,2) &= \phi\Big(\frac{1}{b}\sum_{i=1}^{b} \mathrm{E}[\gamma(\frac{1}{d_i}\|\mathbf{X}_{1i} - \mathbf{X}_{3i}\|^2)]\Big)\\ &= \phi\Big(\frac{1}{b}\sum_{i=1}^{b} \mathrm{E}[\gamma(\frac{1}{d_1}\|\mathbf{X}_{11} - \mathbf{X}_{31}\|^2)]\Big)\\ &= \phi\Big(\mathrm{E}[\gamma(\frac{1}{d_1}\|\mathbf{X}_{11} - \mathbf{X}_{31}\|^2)]\Big),\end{aligned}$$

which implies that $\tilde{h}_b(1,2)$ is free of b. Similarly, we can show that $\tilde{h}_b(1,1) = \phi\big(\mathrm{E}[\gamma(\frac{1}{d_1}\|\mathbf{X}_{11} - \mathbf{X}_{21}\|^2)]\big)$ and $\tilde{h}_b(2,2) = \phi\big(\mathrm{E}[\gamma(\frac{1}{d_1}\|\mathbf{X}_{31} - \mathbf{X}_{41}\|^2)]\big)$ are also free of b. Consequently,
$\liminf_b \tilde{\xi}_b(1,2) (= \tilde{\xi}_1(1,2)$, say) and $\liminf_b \tilde{\tau}_b(1,2) (= \tilde{\tau}_1(1,2)$, say) remain constant for varying b. Clearly, under such circumstances, a sufficient condition for assumption ($A8$) is

$$|\tilde{\xi}_1(1,2) - \tilde{\tau}_1(1,2)| > 0.$$

It is also straightforward to observe that if $\mathrm{E}[\gamma(d_i^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2)] = \mathrm{E}[\gamma(d_{i'}^{-1}\|\mathbf{U}_{i'} - \mathbf{V}_{i'}\|^2)]$ for all $1 \leq i, i' \leq b$, with $\mathbf{U} \sim \mathbf{F}_j$ and $\mathbf{V} \sim \mathbf{F}_{j'}$, then both $\tilde{\xi}_b(j,j')$ and $\tilde{\tau}_b(j,j')$ are also free of b. ■

Lemma 3 *Suppose* $\mathbf{U} = \{\mathbf{U}_1, \mathbf{U}_2, \ldots\}$ *and* $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2, \ldots\}$ *with* $\mathbf{U}_i$ *and* $\mathbf{V}_i$ *denoting the respective sub-vectors for* $i \in \mathbb{N}$. *If* $\mathbf{U}$ *and* $\mathbf{V}$ *are* ρ*-mixing sequences, then the sequence* $\mathbf{W} = (W_1, W_2, \ldots)^\top$, *where* $W_i = \gamma({d_i}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2)$, *is* ρ*-mixing and* $\sum\sum_{1\leq i<i'\leq b} \mathrm{Corr}\,(W_i, W_{i'}) = o(b^2)$.

Proof of Lemma 3 For a random sequence $\mathbf{X} = (X_1, X_2, \ldots)^\top$ we have

$$\rho_{\mathbf{X}}(d) = \sup_{k\geq 1} \rho\big(\sigma(X_1, \ldots, X_k), \sigma(X_{k+d}, \ldots)\big),$$

where $\sigma(X_i, i \in I)$ denotes the σ-field generated by $\{X_i, i \in I\}$, and $\rho(\mathcal{A}, \mathcal{B})$ is defined as $\sup_{X\in\mathscr{L}^2(\mathcal{A}), Y\in\mathscr{L}^2(\mathcal{B})} \big|\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]\big|$. Here, $\mathscr{L}^2(\mathcal{A})$ is the space of square integrable random variables on $\mathcal{A}$. The sequence $\mathbf{X}$ is said to be ρ-mixing if $\rho_{\mathbf{X}}(d) \to 0$ as $d \to \infty$ (see, e.g., Bradley, 2007).

Define $Z_i = h(U_i, V_i)$ for $i \in \mathbb{N}$, where $h : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function. Note that $\sigma(Z_{a_1}, \ldots, Z_{a_2}) \subseteq \sigma(U_{a_1}, \ldots, U_{a_2}) \vee \sigma(V_{a_1}, \ldots, V_{a_2})$. Bradley (2007) showed that

$$\begin{aligned}\rho_{\mathbf{Z}}(d) &= \sup_{k\geq 1} \rho\big(\sigma(Z_1, \ldots, Z_k), \sigma(Z_{k+d}, \ldots)\big) \qquad \text{(A.10)}\\ &\leq \sup_{k\geq 1} \rho\big(\sigma(U_1, \ldots, U_k) \vee \sigma(V_1, \ldots, V_k), \sigma(U_{k+d}, \ldots) \vee \sigma(V_{k+d}, \ldots)\big)\end{aligned}$$

$$
\begin{aligned}
&\text{(see Theorem 3.15-Remark (I), p.82 of Bradley, 2007)}\\
&= \sup_{k\geq 1} \max \left\{\rho\big(\sigma(U_1,\ldots,U_k),\sigma(U_{k+d},\ldots)\big),\ \rho\big(\sigma(V_1,\ldots,V_k),\sigma(V_{k+d},\ldots)\big)\right\}\\
&\text{(see Theorem 6.6-(II) and Note 3, pp.199-200 of Bradley, 2007)}\\
&= \max \left\{\sup_{k\geq 1}\rho\big(\sigma(U_1,\ldots,U_k),\sigma(U_{k+d},\ldots)\big),\ \sup_{k\geq 1}\rho\big(\sigma(V_1,\ldots,V_k),\sigma(V_{k+d},\ldots)\big)\right\}\\
&= \max\left\{\rho_{\mathbf{U}}(d),\rho_{\mathbf{V}}(d)\right\}. \qquad \text{(A.11)}
\end{aligned}
$$

Therefore, $\rho_{\mathbf{Z}}(d) \to 0$ if both $\rho_{\mathbf{U}}(d) \to 0$ and $\rho_{\mathbf{V}}(d) \to 0$ as $d \to \infty$.

Let us consider the sequence $\mathbf{W}$ with $W_1 = g_1(Z_1,\ldots,Z_{d_1})$, $W_2 = g_2(Z_{d_1+1},\ldots,Z_{d_1+d_2})$ and so on, where $g_i : \mathbb{R}^{d_i} \to \mathbb{R}$ for $i \in \mathbb{N}$ are continuous functions. For simplicity, let us assume that $d_i = d_0$ for all $1 \leq i \leq b$. Now, we have

$$
\begin{aligned}
\sigma(W_{a_1},\ldots,W_{a_2}) &= \sigma(g_{a_1}(Z_{(a_1-1)d_0+1},\ldots,Z_{a_1d_0}),\ldots,g_{a_2}(Z_{(a_2-1)d_0+1},\ldots,Z_{a_2d_0}))\\
&\subseteq \sigma(Z_{(a_1-1)d_0+1},\ldots,Z_{a_1d_0},\ldots,Z_{(a_2-1)d_0+1},\ldots,Z_{(a_2-1)d_0}).
\end{aligned}
$$

This further implies that

$$
\begin{aligned}
\rho_{\mathbf{W}}(d) &= \sup_{k\geq 1}\rho\big(\sigma(W_1,\ldots,W_k),\sigma(W_{k+d},\ldots)\big) \qquad \text{(A.12)}\\
&\leq \sup_{k\geq 1}\rho\big(\sigma(Z_1,\ldots,Z_{d_0},\ldots,Z_{(k-1)d_0+1},\ldots,Z_{kd_0}),\sigma(Z_{(k+d-1)d_0+1},\ldots,Z_{(k+d)d_0},\ldots)\big)\\
&\quad \text{(see Theorem 3.15-Remark (I), p.82 of Bradley, 2007)}\\
&\leq \sup_{k\geq 1}\rho\big(\sigma(Z_1,\ldots,Z_{d_0},\ldots,Z_{(k-1)d_0+1},\ldots,Z_{kd_0}),\sigma(Z_{kd_0+d},Z_{kd_0+d+1}\ldots)\big)\\
&= \sup_{k\geq 1}\rho\big(\sigma(Z_1,\ldots,Z_k),\sigma(Z_{k+d},\ldots)\big)\\
&= \rho_{\mathbf{Z}}(d). \qquad \text{(A.13)}
\end{aligned}
$$

Proof for the case when d_is are *unequal, but bounded* follows by using a similar line of arguments. From equations (A.10) and (A.12), it follows that $\mathbf{W}$ is a ρ-mixing sequence if both the original sequences $\mathbf{U}$ and $\mathbf{V}$ are ρ-mixing. Consider the maps $h(u,v) = (u-v)^2$, $g(u_1,\ldots,u_k) = (u_1+\cdots+u_k)/k$, and γ as described in Lemma 2.3. Hence, if $\mathbf{U}$ and $\mathbf{V}$ are ρ-mixing, then the sequence $\mathbf{W} = \{W_i = \gamma({d_i}^{-1}\|\mathbf{U}_i - \mathbf{V}_i\|^2), i \geq 1\}$ is also ρ-mixing.

Now, by Theorem 4.5(b) of Bradley (2007), we have

$$
\mathrm{Corr}(W_i,W_{i'}) \leq \rho(\sigma(W_i),\sigma(W_{i'})) \leq \rho\big(\sigma(W_1,\ldots,W_i),\sigma(W_{i'},\ldots)\big) \leq \rho_{\mathbf{W}}(i'-i).
$$

Therefore,

$$
0 \leq b^{-2}\sum\sum_{1\leq i<i'\leq b}\mathrm{Corr}(W_i,W_{i'}) \leq b^{-2}\sum\sum_{1\leq i<i'\leq b}\rho_{\mathbf{W}}(i'-i) \leq b^{-2}\sum_{l=1}^{b}(b-l)\rho_{\mathbf{W}}(l) \leq b^{-1}\sum_{l=1}^{b}\rho_{\mathbf{W}}(l).
$$

Since, $\rho_{\mathbf{W}}(b) \to 0$ as $b \to \infty$, it follows from Cesàro summability that

$$
\sum\sum_{1\leq i<i'\leq b}\mathrm{Corr}\,(W_i,W_{i'}) = o(b^2).
$$

■

Appendix B. Notations

Table 4: List of standard notations

Symbol	*Denotes*
J	number of classes
n_j	training sample size of j-th class
n	total training sample size
d	data dimension
$\mathscr{X}$	random sample
$\boldsymbol{\mu}$	location parameter (vector)
Σ	scale parameter (matrix)
ρ	population correlation coefficient
r	sample correlation coefficient
X	random variable
$\mathbf{X}$	random vector
F	distribution function of a random variable X
$\mathbf{F}$	distribution function of a random vector $\mathbf{X}$
$\mathrm{Corr}(X, Y)$	correlation between X and Y
δ	a generic classifier
Δ	misclassification probability (rate) of the classifier δ

Table 5: Notations specific to this paper

Symbol	*Denotes*	*Remark*
b	number of blocks	
$h(\mathbf{U}, \mathbf{V})$	generalized distance between $\mathbf{U}$ and $\mathbf{V}$	
$\xi(\mathbf{U}, \mathbf{V})$	measure of dissimilarity between $\mathbf{U}$ and $\mathbf{V}$	average distance classifier
$\tilde{\xi}(j, j')$	measure of separability between class j and j'	average distance classifier
$\tau(\mathbf{U}, \mathbf{V})$	measure of dissimilarity between $\mathbf{U}$ and $\mathbf{V}$	nearest neighbor classifier
$\tilde{\tau}(j, j')$	measure of separability between class j and j'	nearest neighbor classifier

References

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

Makoto Aoshima, Dan Shen, Haipeng Shen, Kazuyoshi Yata, Yi-Hui Zhou, and James S Marron. A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, 60(1):4–19, 2018.

L. Baringhaus and C. Franz. Rigid motion invariant two-sample tests. *Statistica Sinica*, 20 (4):1333–1361, 2010.

Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

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

Richard C Bradley. *Introduction to Strong Mixing Conditions.* Heber City: Kendrick Press, 2007.

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

Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The UCR time series classification archive, October 2018. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

Sampath Deegalla and Henrik Bostrom. 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, 2006.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

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

Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

Daniel P Faith and Paul A Walker. Environmental diversity: On the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation*, 5(4):399–415, 1996.

William Feller. *An Introduction to Probability Theory and its Applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

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

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* Springer, New York, 2009.

Debasis Kundu and Rameshwar D. Gupta. Power-normal distribution. *Statistics*, 47(1): 110–125, 2013.

Arnab K Pal, Pronoy K Mondal, and Anil K Ghosh. High dimensional nearest neighbor classification based on mean absolute differences of inter-point distances. *Pattern Recognition Letters*, 74:1–8, 2016.

Soham Sarkar and Anil K Ghosh. On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat*, 7(1):e187, 2018.

Soham Sarkar, Rahul Biswas, and Anil K Ghosh. On high-dimensional modifications of some graph-based two-sample tests. *Machine Learning*, 109:279–306, 2020.

Gábor J Székely and Maria L Rizzo. The energy of data. *Annual Review of Statistics and Its Application*, 4:447–479, 2017.

Vladimir Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.