Binary Discrimination Methods for High Dimensional Data with a Geometric Representation

Addy Bolivar-Cime*, Luis Miguel Cordova-Rodriguez

Universidad Juárez Autónoma de Tabasco,

División Académica de Ciencias Básicas

Received: November 11, 2016

Abstract

Four binary discrimination methods are studied in the context of HDLSS data with an asymptotic geometric representation, when the dimension increases while the sample sizes of the classes are fixed. We show that the methods Support Vector Machine, Mean Difference, Distance Weighted Discrimination and Maximal Data Piling have the same asymptotic behavior as the dimension increases. We study the consistent, inconsistent and strongly inconsistent cases in terms of angles between the normal vectors of the separating hyperplanes of the methods and the optimal direction for classification. A simulation study is done to assess the theoretical results.

1 Introduction

The asymptotic geometric representation of multivariate data as the dimension increases while the sample size is fixed, has been studied by Ahn et al. (2007), Hall et al. (2005) and Qiao et al. (2010). They show conditions under which the High dimension, low sample size (HDLSS) data tend to lie deterministically at the vertices of a regular simplex, similarly to the multivariate standard Gaussian data. This geometric structure is used to analyze the behavior of some statistical methodologies for multivariate data in the HDLSS setting. In particular, in Ahn et al. (2007), Jung and Marron (2009) and Jung et al. (2012) it is discussed the behavior of Principal Component Analysis under the geometric representation of HDLSS data. In Hall et al. (2005) it is studied, in terms of probability of misclassification, the behavior of some

*Corresponding author. División Académica de Ciencias Básicas - UJAT, Carretera Cunduacán-Jalpa KM. 1, Col. La Esmeralda, CP. 86690. E-mail: addy.bolivar@ujat.mx.

binary discrimination methods, including the methods Support Vector Machine (Cristianini and Shawe-Taylor (2000), Vapnik (1995)), Distance Weighted Discrimination (Marron (2015), Marron et al. (2007)) and Mean Difference (Scholkopf and Smola (2002)), when the data have this asymptotic geometric representation as the dimension increases. In Qiao et al. (2010) a similar analysis is done for the binary discrimination method Weighted Distance Weighted Discrimination (wDWD), they also study the asymptotic behavior of this method in terms of the angle between the normal vector of the separating hyperplane and the optimal direction for classification.

In Bolivar-Cime and Marron (2013), considering Gaussian data with common diagonal covariance matrix, it is shown that the four methods Support Vector Machine (SVM), Distance Weighted Discrimination (DWD), Mean Difference (MD) and Maximal Data Piling (MDP) (Ahn and Marron (2010)) have the same asymptotic behavior as the dimension increases and the sample sizes are fixed, in terms of angles between the normal vectors of the separating hyperplanes of the methods. In the present paper we prove that this result of Bolivar-Cime and Marron (2013) holds for more general HDLSS data with an asymptotic geometric representation. Note that due to the asymptotic geometric representation of the data, if the two classes of the training data set have the same distribution except that one class has mean v_d and the other class has mean zero, the vector v_d is the optimal direction for the normal vector of a separating hyperplane of the data. We showed that as the dimension d increases the angles between the normal vectors of the separating hyperplanes of the four methods and v_d converge to zero in probability when $\| v_d \| \gg d^{1/2}$, i.e. are *consistent*; and converge to $\pi/2$ in probability when $\| v_d \| \ll d^{1/2}$, i.e. are *strongly inconsistent*. In the case where $\| v_d \| \approx cd^{1/2}$ with $0 < c < \infty$, we showed that the angles converge to a number in the interval $(0, \pi/2)$ in probability as the dimension increases, i.e. the four methods are *inconsistent*. We provide some examples of HDLSS data for which our results are valid.

The results of the present paper complement the results of Hall et al. (2005) about the behavior of some binary classification methods. Furthermore, our results extend the results of Ahn et al. (2007) and Qiao et al. (2010), since in Ahn et al. (2007) the method MDP is not considered, in Qiao et al. (2010) only the methods DWD and wDWD are considered. Our results also provide a theoretical explanation of the phenomena observed in the simulations studies of Ahn and Marron (2010) and Marron et al. (2007), in terms of means of misclassification rates. Additionally, we compare the asymptotic behavior of the angles between the normal vectors of the four methods and the optimal direction with the asymptotic behavior of the probabilities of misclassification. We also present a simulation study to numerically assess our theoretical results.

As is mentioned in Hall et al. (2005), when $d \geq N$, where N is the sample size of the data, and no k data points lie in a $(k-2)$-dimensional hyperplane (which happens with probability one for data with continuous

probability densities), the training data set is linearly separable. In this paper we restrict our attention to the linearly separable case, assuming that the HDLSS data set treated here is the linearly separable with probability one.

This paper is divided as follows. In Section 2 we present the geometric representation of some HDLSS data. Our theoretical results about the asymptotic behavior of the normal vectors of the separating hyperplanes of the four methods are presented in Section 3. In Section 4 we present the asymptotic behavior of the probabilities of misclassification of the four methods. In Section 5 we provide a simulation study to evaluate our results. The technical details of the paper are presented in Section 6. Finally, we provide conclusions in Section 7.

2 Geometric representation of high dimensional data

The geometric representation of high dimensional data concerns the geometric structure that multivariate data have as the dimension tends to infinity while the sample size is fixed. This geometric structure can be found for example in multivariate standard Gaussian data as the dimension tends to infinity.

2.1 Standard Gaussian geometrical representation

As it is mentioned in Hall et al. (2005), if Z has d-multivariate standard Gaussian distribution, as d tends to infinity we have that

$$\| Z \| = d^{1/2} + O_p(1). \tag{1}$$

This means that as the dimension increases the random vector Z tends to lie near the surface of an expanding sphere. Furthermore, if Z_1 and Z_2 are two independent d-multivariate standard Gaussian vectors, as d increases we have

$$\| Z_1 - Z_2 \| = (2d)^{1/2} + O_p(1). \tag{2}$$

Thus, the distance between the data vectors is approximately constant as the dimension increases. It is also true that for these vectors, as the dimension increases we have

$$\text{Angle}(Z_1, Z_2) = \frac{\pi}{2} + O_p(d^{-1/2}). \tag{3}$$

That is, the angle between the vectors tends to be an orthogonal angle as the dimension tends to infinity.

In general, if we have n of these independent d-multivariate standard Gaussian vectors, as the dimension increases, all pairwise distances are approximately equal and all pairwise angles are approximately perpendicular. Because all pairwise distances are nearly the same, the n vectors tend to be the vertices of a regular n-polyhedron, that is a polyhedron with n vertices and with edges of the same length. This n-polyhedron is called an *n-simplex*.

2.2 General geometrical representation

In Ahn et al. (2007), Hall et al. (2005) and Qiao et al. (2010) it is shown that the approximate n-simplex structure of the standard Gaussian data can be observed for more general data, as it is presented below.

Let $X(d) = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})^\top$ be the vector obtaining by truncating an infinite time series, which is written as the vector $X = (X^{(1)}, X^{(2)}, \ldots)^\top$. Let $\mathcal{X}(d) = \{X_1(d), X_2(d), \ldots, X_n(d)\}$ be a random sample of independent and identically distributed random vectors with the same distribution as $X(d)$. Assume the following:

(a) The fourth moments of the entries of the data vectors are uniformly bounded.

(b) For a constant σ^2,

$$\frac{1}{d}\sum_{k=1}^{d} \mathrm{var}(X^{(k)}) \longrightarrow \sigma^2 \qquad \text{as } d \to \infty. \tag{4}$$

(c) The time series X is *ρ-mixing for functions that are dominated by quadratics*, in the sense that whenever functions f and g of two variables satisfy $|f(u,v)| + |g(u,v)| \le Cu^2v^2$ for fixed $C > 0$ and all u and v, we have

$$\sup_{1\le k,l<\infty, |k-l|\ge r} |\mathrm{corr}[f(U^{(k)}, V^{(k)}), g(U^{(l)}, V^{(l)})]| \le \rho(r), \tag{5}$$

with $(U,V) = (X,X),\ (X,X')$, where X' is independent and has the same distribution as X, and the function ρ satisfies $\rho(r) \to 0$ as $r \to \infty$.

Under the conditions (a), (b) and (c), in Hall et al. (2005) it is shown that the distance between $X_i(d)$ and $X_j(d)$, for $i \neq j$, is approximately $(2\sigma^2 d)^{1/2}$ when d is large, in the sense that

$$\frac{\| X_i - X_j \|^2}{d} \xrightarrow{P} 2\sigma^2 \qquad \text{as } d \to \infty, \tag{6}$$

where "$\xrightarrow{P}$" means convergence in probability. It is also true that

$$\frac{\| X_i \|^2}{d} \xrightarrow{P} \sigma^2 \qquad \text{as } d \to \infty. \tag{7}$$

Therefore, the asymptotic n-simplex structure of the standard Gaussian data is also observed for these data.

As it is mentioned in Ahn et al. (2007), condition (c) states that the variables have to be *nearly independent*, since they must satisfy a ρ-mixing condition. However, this condition is too strict because is common to have strong collinearity among variables. Furthermore, this condition depends on the order of the data entries, which can be arbitrary in many applications. In Ahn et al. (2007) and Qiao et al. (2010) it is shown that the asymptotic n-simplex structure of high dimensional data can be observed under mild conditions, which are presented below.

Let $\mathbf{X}_d = [X_1, X_2, \ldots, X_n]$ be a $d \times n$ data matrix with $d > n$, where the random vectors $X_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)})^\top$, $i = 1, 2, \ldots, n$, are independent and identically distributed from a d-dimensional multivariate distribution with mean zero and nonnegative definite covariance matrix Σ_d. Suppose that the eigenvalue decomposition of Σ_d is $\Sigma_d = V_d \Lambda_d V_d^\top$, where Λ_d is the diagonal matrix of eigenvalues $\lambda_{1,d} \geq \lambda_{2,d} \geq \cdots \geq \lambda_{d,d} \geq 0$ and V_d is the matrix of corresponding eigenvectors. If Σ_d is positive definite (and all its eigenvalues are positive) define $\widetilde{\mathbf{Z}}_d = \Lambda_d^{-1/2} V_d^\top \mathbf{X}_d$, which is a $d \times n$ random data matrix from a distribution with identity covariance matrix. Observe that, if the columns of $\mathbf{X}_d$ are Gaussian, the elements of $\widetilde{\mathbf{Z}}_d$ are independent standard Gaussian univariate variables.

The *sample covariance matrix* is given by $S_d = n^{-1} \mathbf{X}_d \mathbf{X}_d^\top$, because the population mean is the zero vector. The *dual sample covariance matrix* is defined as $S_{D,d} = n^{-1} \mathbf{X}_d^\top \mathbf{X}_d$, which is an $n \times n$ matrix. It is important to note that $S_{D,d}$ has the same nonzero eigenvalues as S_d. Using the fact that $V_d^\top V_d$ is the identity, we have

$$n S_{D,d} = \widetilde{\mathbf{Z}}_d^\top \Lambda_d \widetilde{\mathbf{Z}}_d = \sum_{i=1}^{d} \lambda_{i,d} W_{i,d}, \tag{8}$$

where $W_{i,d} = \widetilde{Z}_{i,d}^\top \widetilde{Z}_{i,d}$ and $\widetilde{Z}_{i,d}$, for $i = 1, 2, \ldots, d$, are the row vectors of $\widetilde{\mathbf{Z}}_d$. Note that if $\mathbf{X}_d$ is Gaussian, $W_{i,d}$, for $i = 1, 2, \ldots, d$, are independent matrices from the Wishart distribution $\mathcal{W}_n(1, \mathrm{I}_n)$. Assume the following for the matrix $\mathbf{X}_d$:

(a') The fourth moments of the variables are uniformly bounded.

(b') The representation in (8) holds.

(c') The eigenvalues of Σ_d are sufficiently diffused, in the sense that

$$\frac{\sum_{i=1}^{d} \lambda_{i,d}^2}{(\sum_{i=1}^{d} \lambda_{i,d})^2} \longrightarrow 0 \qquad \text{as } d \to \infty. \tag{9}$$

(d') The entries of $\widetilde{\mathbf{Z}}_d$ are independent.

In Ahn et al. (2007) and Qiao et al. (2010) it is shown that under the conditions (a')–(d'), the square distance between X_i and X_j, for $i \neq j$, is approximately $2\sum_{i=1}^{d} \lambda_{i,d}$ when d is large, in the sense that

$$\frac{\| X_i - X_j \|^2}{\sum_{i=1}^{d} \lambda_{i,d}} \xrightarrow{P} 2 \qquad \text{as } d \to \infty. \tag{10}$$

It is also true that

$$\frac{\| X_i \|^2}{\sum_{i=1}^{d} \lambda_{i,d}} \xrightarrow{P} 1 \qquad \text{as } d \to \infty. \tag{11}$$

Therefore the data tend to form an n-simplex as dimension increases. It is also shown in Ahn et al. (2007) that the condition (c') is milder than the ρ-mixing condition (c). In Jung and Marron (2009) it is shown that condition (d') can be relaxed assuming that the entries of $\widetilde{\mathbf{Z}}_d$ are ρ-mixing under some permutation, however this last condition is still very strict. By the results of Yata and Aoshima (2012) we have (10) and (11) under the conditions (a')–(c') and the new condition

$$\frac{\sum_{s,t=1}^{d} \lambda_{s,d}\lambda_{t,d} E\{(Z_{1(s)}^2 - 1)(Z_{1(t)}^2 - 1)\}}{\mathrm{tr}(\Sigma_d)^2} \longrightarrow 0 \quad \text{as } d \to \infty, \tag{12}$$

where $Z_{1(s)}$ is the first element of $\widetilde{Z}_{s,d}$. In Yata and Aoshima (2012) it is mentioned that (12) is milder than (d') (or the ρ-mixing condition for the entries of $\widetilde{\mathbf{Z}}_d$), since (12) holds under (c') and (d') (or the ρ-mixing condition for the entries of $\widetilde{\mathbf{Z}}_d$).

3 Asymptotic behavior of the normal vectors

In this section we present a generalization of the Theorem 3.1 of Bolivar-Cime and Marron (2013). That theorem states that the asymptotic behavior of the four binary discrimination methods SVM, DWD, MD and MDP is the same as the dimension increases, when the two classes C_+ and C_- are Gaussian with means v_d and zero, respectively, and common diagonal covariance matrix $\Sigma_d = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, where $\{\sigma_k\}_{k=1}^{\infty}$ is a bounded sequence of positive numbers such that $\sum_{k=1}^{d} \sigma_k^2/d \longrightarrow \sigma^2$ as $d \to \infty$, for some $\sigma > 0$. It can be

seen that these data have the asymptotic geometric representation of Section 2.2, and since the difference between the two classes is determined by the mean vector v_d, the optimal direction for the normal vector of a separating hyperplane of these data is v_d. Specifically, the Theorem 3.1 of Bolivar-Cime and Marron (2013) states that, under the above assumptions, when $\parallel v_d \parallel d^{-1/2} \to \infty$ the angles between the normal vectors of the separating hyperplanes of the four methods and the optimal direction v_d converge to zero in probability as $d \to \infty$, i.e. are consistent; when $\parallel v_d \parallel d^{-1/2} \to 0$ these angles converge to $\pi/2$ in probability, i.e. are strongly inconsistent; and when $\parallel v_d \parallel d^{-1/2} \to c$ with $0 < c < \infty$, these angles converge to $\arccos(c/(\gamma\sigma^2 + c^2)^{1/2})$, where $\gamma = \frac{1}{m} + \frac{1}{n}$ with m and n the sample sizes of C_+ and C_-, respectively, i.e. are inconsistent.

Our next theorem claims that when we consider multivariate data with an asymptotic geometric representation, similar to that of the multivariate standard Gaussian data or the multivariate data of Section 2.2, the result of Theorem 3.1 of Bolivar-Cime and Marron (2013) still holds under some conditions.

Theorem 3.1 *Let m, n be positive integers and let $N = m + n$. Let $Z_1, Z_2, \ldots, Z_N$ be independent and identically distributed d-dimensional random vectors, with mean zero and covariance matrix Σ_d. Let C_+ be the class of the random vectors $X_i = Z_i + v_d$, for $i = 1, 2, \ldots, m$, and let C_- be the class of the random vectors $Y_j = Z_{m+j}$, $j = 1, 2, \ldots, n$, where $\parallel v_d \parallel d^{-1/2} \to c$, with $0 \leq c \leq \infty$. Assume the following:*

(i) The random vectors have the asymptotic geometric representation

$$\frac{\parallel Z_i \parallel^2}{d} \xrightarrow{P} \sigma^2 \qquad \text{and} \qquad \frac{\parallel Z_i - Z_j \parallel^2}{d} \xrightarrow{P} 2\sigma^2 \tag{13}$$

as $d \to \infty$ for some $\sigma > 0$, for all $i, j = 1, 2, \ldots, N$ and $i \neq j$.

(ii) The covariance matrix $\Sigma_d = (\sigma_{i,j})$ and the vector $v_d = (v_d^{(1)}, v_d^{(2)}, \ldots, v_d^{(d)})^\top$ satisfy

$$\frac{D_{\Sigma_d}(v_d, 0)^2}{d \parallel v_d \parallel^2} = \frac{\sum_{k,r=1}^{d} \sigma_{k,r} v_d^{(k)} v_d^{(r)}}{d \parallel v_d \parallel^2} \longrightarrow 0 \qquad \text{as } d \to \infty, \tag{14}$$

where $D_{\Sigma_d}(x, y) = [(x - y)^\top \Sigma_d (x - y)]^{1/2}$, $\forall x, y \in \mathbb{R}^d$, is the Mahalanobis distance corresponding to Σ_d.

Under these conditions if v represents the normal vector of the MD, SVM, DWD or MDP hyperplane of the

training data set, then

$$\text{Angle}(v, v_d) \xrightarrow{P} \begin{cases} 0, & \text{if } c = \infty; \\ \dfrac{\pi}{2}, & \text{if } c = 0; \\ \arccos\left(\dfrac{c}{(\gamma\sigma^2 + c^2)^{1/2}}\right), & \text{if } 0 < c < \infty; \end{cases}$$

as $d \to \infty$, *where* $\gamma = \frac{1}{m} + \frac{1}{n}$.

As in Bolivar-Cime and Marron (2013), we observe in our Theorem 3.1 that the asymptotic behavior of the normal vectors of the four methods is related with the distance between the two classes, in particular with $\| v_d \|$. It is observed that when $\| v_d \| \gg d^{1/2}$ $(c = \infty)$ the classification is easier than when $\| v_d \| \ll d^{1/2}$ $(c = 0)$. This is explained by the geometric representation of the data sets, since the data tend to lie at a distance $\sigma d^{1/2}$ from the mean when d is large.

To illustrate the role of c in the last theorem, we present in the Figure 1 the intuitive idea of the asymptotic behavior of the data, with $\sigma = 1$ and $0 < c < \infty$. By the asymptotic geometric representation of the data, when dimension is large the data of the class C_- will be around the sphere of radio $d^{1/2}$ with center in the origin, while the data of the class C_+ will be around the sphere of radio $d^{1/2}$ and center v_d. We also have that $\| v_d \| \approx cd^{1/2}$. Therefore, as c approximates to infinity the two spheres are far away and the classification with the four methods is easier, however as c approximates to zero the two spheres are very close and the classification is more difficult.

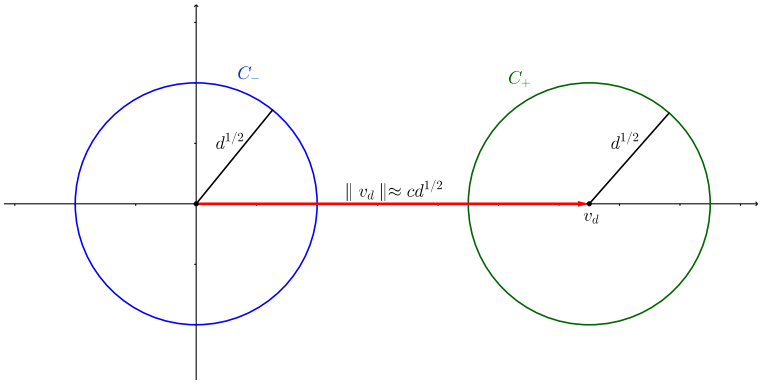


Figure 1: Asymptotic behavior of the data, with $\sigma = 1$ and $0 < c < \infty$. The classification is easier when the two spheres are far away $(c \to \infty)$ and it is more difficult when the two spheres are very close $(c \to 0)$.

Observe that condition (14) is in terms of a Mahalanobis distance between the two class means. Furthermore, since $\Sigma_d = \Sigma_d^{1/2}\Sigma_d^{1/2}$, where $\Sigma_d^{1/2} = V\Lambda^{1/2}V^\top$, with Λ the diagonal matrix of eigenvalues of Σ_d and V the orthogonal matrix of corresponding eigenvectors, we have that $D_{\Sigma_d}(x, y) = \| \Sigma_d^{1/2}x - \Sigma_d^{1/2}y \|$, i.e.

$D_{\Sigma_d}(x, y)$ is the euclidean distance between the vectors obtained from x and y by using the linear transformation $\Sigma_d^{1/2}$. Thus, $D_{\Sigma_d}(v_d, 0) = \parallel \Sigma_d^{1/2} v_d \parallel$ is the euclidean distance between that linear transformation of the class means. Hence, (14) is equivalent to $\parallel \Sigma_d^{1/2} v_d \parallel^2 / (d \parallel v_d \parallel^2) \to 0$ as $d \to \infty$.

We also have that condition (14) is equivalent to $D_{\Sigma_d}(v_d, 0) / \parallel v_d \parallel = o(d^{1/2})$ as $d \to \infty$, which is satisfied in particular if the ratio $D_{\Sigma_d}(v_d, 0) / \parallel v_d \parallel$ is bounded. Therefore, condition (14) tries to control the magnitude of the Mahalanobis distance with respect to the euclidean distance between the two class means. For example, if $v_d = (cd^{1/2}, 0, \ldots, 0)^\top$ with $0 < c < \infty$, and

$$\Sigma_d = \begin{bmatrix} 2\mathrm{I}_{d/2} & \mathbf{0} \\ \mathbf{0} & \mathrm{I}_{d/2} \end{bmatrix} \tag{15}$$

with d even, then $D_{\Sigma_d}(v_d, 0) / \parallel v_d \parallel = 2^{1/2}$ and condition (14) is satisfied. For this example we observe that the Mahalanobis distance between the two class means is $2^{1/2}$ times the euclidean distance between them.

Remark 3.1 *The condition (i) of the last theorem is satisfied if conditions (a)–(c) of Section 2.2 hold. If $\lambda_{1,d} \geq \cdots \geq \lambda_{d,d}$ are the eigenvalues of Σ_d and $\sum_{i=1}^{d} \lambda_{i,d}/d \longrightarrow \sigma^2$ as $d \to \infty$, for some $\sigma > 0$, then condition (i) is also satisfied under the conditions (a')–(d') of Section 2.2 or the conditions (a')–(c') and (12), because of (10) and (11).*

Remark 3.2 *There are several cases where the condition (ii) is satisfied, some of them are the following (see Section 6 for details):*

(I) The covariance matrix Σ_d is a diagonal matrix with entries uniformly bounded.

(II) The vector v_d has a fixed number of nonzero entries, and the second moments of the entries of Z_1 are uniformly bounded.

(III) The entries of v_d are uniformly bounded, $\parallel v_d \parallel d^{-1/2} \to c$ with $0 < c < \infty$, and one of the following conditions is satisfied:

a) the entries of Z_1 have second moments uniformly bounded, and are ρ-mixing in the sense that

$$\sup_{|k-l| \geq r} |E(Z_1^{(k)} Z_1^{(l)})| = \sup_{|k-l| \geq r} |\sigma_{k,l}| = \rho(r) \longrightarrow 0 \qquad \text{as } r \to \infty.$$

b) the entries of Z_1 have second moments uniformly bounded, and Σ_d has a fixed number of nonzero upper diagonals.

c) the eigenvalues of Σ_d, $\lambda_{1,d} \geq \lambda_{2,d} \geq \cdots \geq \lambda_{d,d} \geq 0$, satisfy

$$\frac{\sum_{k=1}^{d} \lambda_{k,d}^2}{d^2} \longrightarrow 0 \qquad \text{as } d \to \infty. \tag{16}$$

(IV) The vector v_d has the form $v_d = \beta \mathbf{1}_d$, with $\beta = \beta_d \to c$ as $d \to \infty$, where $0 \leq c \leq \infty$, and one of the conditions a), b) or c) of (III) is satisfied.

By Remark 3.1, the multivariate Gaussian data of Theorem 3.1 of Bolivar-Cime and Marron (2013) satisfy the condition (i) of Theorem 3.1. These data also satisfy the condition (I) of Remark 3.2, therefore the condition (ii) of Theorem 3.1 is satisfied as well. In this sense Theorem 3.1 generalizes Theorem 3.1 of Bolivar-Cime and Marron (2013), by extending this result to more general multivariate data with an asymptotic geometric representation. Theorem 3.1 also provide a theoretical explanation of the simulation results presented in Ahn and Marron (2010) and Marron et al. (2007), where it is observed that some of the considered binary discrimination methods have approximately the same behavior, in terms of means of error rates (misclassification rates), as the dimension increases.

For the proof of Theorem 3.1 we need the next lemma, which is also a generalization of a result in Bolivar-Cime and Marron (2013). As it is explained in Bolivar-Cime and Marron (2013), the normal vectors of the MD, SVM and DWD hyperplanes are proportional to the difference between two points on the convex hulls of the two classes. The next lemma provides an explicit asymptotic representation for these differences, when $0 \leq c < \infty$ (the inconsistent cases). We denote by $\alpha = (\alpha_+^\top, \alpha_-^\top)^\top$ an N-dimensional vector, where $\alpha_+ = (\alpha_{1+}, \alpha_{2+}, \ldots, \alpha_{m+})^\top$ and $\alpha_- = (\alpha_{1-}, \alpha_{2-}, \ldots, \alpha_{n-})^\top$ are subvectors of α of dimensions m and n, respectively.

Lemma 3.1 *Assume the same as in Theorem 3.1. Suppose that* $\parallel v_d \parallel d^{-1/2} \to c$ *with* $0 \leq c < \infty$. *Let* $\mathbf{X} = [X_1, X_2, \ldots, X_m]$ *and* $\mathbf{Y} = [Y_1, Y_2 \ldots, Y_n]$. *If the vector* $\widetilde{v} = \mathbf{X}\alpha_+ - \mathbf{Y}\alpha_-$, *with* $\alpha \geq \mathbf{0}$ *and* $\mathbf{1}_m^\top \alpha_+ = \mathbf{1}_n^\top \alpha_- = 1$, *is proportional to the normal vector of the MD, SVM or DWD hyperplane we have that*

$$\alpha_{i+} \xrightarrow{P} \frac{1}{m}, \quad \alpha_{j-} \xrightarrow{P} \frac{1}{n}, \tag{17}$$

as $d \to \infty$, *for* $i = 1, 2, \ldots, m$ *and* $j = 1, 2, \ldots, n$.

Thus, in the inconsistent cases the normal vectors of SVM and DWD are approximately in the same direction as the normal vector of MD when d is large. This is also true for MDP, as we will see in the proof of Theorem 3.1. Due to the asymptotic geometric representation of the data (see Figure 1), in Theorem 3.1 we have that as $c \to \infty$ the angles between the normal vectors of the four methods and v_d tend to zero as d

increases, that is, in this case the direction of the four methods is approximately the direction of v_d when d is large.

4 Asymptotic properties of the probabilities of misclassification

Assume the same as in Theorem 3.1. Suppose $\| v_d \| / d^{1/2} \to c$, with $0 < c < \infty$. Due to the asymptotic geometric representation of the data, by Hall et al. (2005) and Qiao and Zhang (2015) we have the following two results about the asymptotic error rates of the SVM, MD and DWD hyperplanes.

Theorem 4.1 *Assume that $n \geq m$; if need be, interchange X and Y to achieve this. If $c > \sigma(1/m - 1/n)^{1/2}$, then the probability that a new datum from either the X-population or the Y-population is correctly classified by the SVM or the MD hyperplane converges to 1 as $d \to \infty$. If $c < \sigma(1/m - 1/n)^{1/2}$, then with probability converging to 1 as $d \to \infty$ a new datum from either population will be classified by the SVM or the MD hyperplane as belonging to the Y-population.*

Theorem 4.2 *Assume that $n \geq m$; if need be, interchange X and Y to achieve this. If $c > \sigma[(n/m)^{1/2}/m - 1/n]^{1/2}$, then the probability that a new datum from either the X-population or the Y-population is correctly classified by the DWD hyperplane converges to 1 as $d \to \infty$. If $c < \sigma[(n/m)^{1/2}/m - 1/n]^{1/2}$, then with probability converging to 1 as $d \to \infty$ a new datum from either population will be classified by the DWD hyperplane as belonging to the Y-population.*

As we mentioned before, under the hypotheses of Theorem 3.1 and if $0 < c < \infty$, the normal vector of MDP is approximately in the same direction as the normal vector of MD when d is large. Therefore, if we take the intercept of the MDP hyperplane as $b = -v^\top(\overline{X} + \overline{Y})/2$, where $\overline{X}$ and $\overline{Y}$ are the class means of the X and Y populations, respectively, and v is the normal vector of MDP, then the MDP hyperplane coincides with the MD hyperplane when d tends to infinity. Thus, Theorem 4.1 holds for the MDP method.

By the above results, if $m = n$ the four methods give asymptotically correct classification of a new datum from any population, for all $0 < c < \infty$. In the case where the sample sizes m and n are unequal, for example if $n > m$, define

$$M_1 = \sigma(1/m - 1/n)^{1/2}, \quad M_2 = \sigma[(n/m)^{1/2}/m - 1/n]^{1/2},$$

and note that $M_2 > M_1$. By the last theorems, if $c > M_2$ the four methods give asymptotically correct classification of a new datum from any population; if $M_2 > c > M_1$ then SVM, MD and MDP give asymptotically correct classification of a new datum from any population, while DWD gives asymptotically perfect classification for the Y-population and asymptotically completely incorrect classification for the

X-population. This shows an asymptotic advantage of SVM, MD and MDP over DWD, in the sense of classifying correctly new data from any population for a wider range of values of c as d tends to infinity.

Observe that if $n \geq m$ and $c > M_2$, by the last results the four methods have the consistent property of the error rates in the sense that their error rates tend to zero as d tends to infinity, and by Theorem 3.1 the four methods have the inconsistent property of the normal vectors in the sense that the angles of their normal vectors and the optimal direction do not tend to zero as d tends to infinity. That is, in this case the asymptotic geometric representation of the data allows to find separating hyperplanes that give perfect classification even when their normal vectors are not in the same direction as the optimal direction. By Theorem 3.1 we have that when $d \to \infty$ the limit of the angles between the normal vectors of the four methods and the optimal direction approaches to zero as c tends to infinity, however it is sufficient to have $c > M_2$ in order to have asymptotically correct classification of a new datum from any population with the four methods, and in this situation the limit of the angles between the normal vectors of the four methods and the optimal direction is at most $\arccos\left(\dfrac{M_2}{(\gamma\sigma^2 + M_2^2)^{1/2}}\right)$.

In the case when n and m are unequal, the above results also show that the classification is easier when $c \to \infty$ than when $c \to 0$. In the case when $n = m$ this is also true, since as we will see in the simulation study of Section 5, even when the four methods give asymptotically correct classification of a new datum from any population for all $0 < c < \infty$, as c increases the convergence of the error rates to zero as d tends to infinity is faster.

Aoshima and Yata (2014) and Nakayama et al. (2017) proposed a bias-corrected MD and a bias-corrected SVM, respectively, to improve the performance of the error rates of MD and SVM. Theorem 3.1 also holds for the bias-corrected classifiers, since it is only about the normal vectors of the separating hyperplanes. Assume the hypotheses of Theorem 3.1. Consider the bias-corrected MD proposed by Aoshima and Yata (2014), named the distance-based classifier, which is defined as follows: One classifies an individual X_0 into C_+ if $W(X_0) < 0$ and into C_- otherwise, where $W(X_0) = (X_0 - (\overline{X} + \overline{Y})/2)^\top (\overline{Y} - \overline{X}) - \mathrm{tr}S_+/(2m) + \mathrm{tr}S_-/(2n)$, S_+ and S_- is the sample covariance matrix for C_+ and C_-, respectively. Here, $-\mathrm{tr}S_+/(2m) + \mathrm{tr}S_-/(2n)$ is a bias-correction term. This classifier is equivalent to the scale adjusted distance-based classifier given by Chan and Hall (2009). From Theorem 1 of Aoshima and Yata (2014), the error rates of $W(X_0)$ tend to zero as $d \to \infty$ if

$$\frac{D_{\Sigma_d}(v_d, 0)^2}{\| v_d \|^4} \longrightarrow 0 \quad \text{and} \quad \frac{\mathrm{tr}(\Sigma_d^2)}{\min(m, n) \| v_d \|^4} \longrightarrow 0, \quad \text{as } d \to \infty. \tag{18}$$

As a referee pointed out, if (I) in Remark 3.2 holds, it holds that $\mathrm{tr}(\Sigma_d^2) = O(d)$ and $D_{\Sigma_d}(v_d, 0)^2 = O(\| v_d \|^2)$. Therefore, if $\| v_d \|^2 / d \to c^2 > 0$ then (18) holds and the error rates of the classifier $W(X_0)$ tend to zero

as $d \to \infty$, even when the angle between the normal vector of the separating hyperplane and the optimal direction do not tend to zero. Observe that (18) holds even when $\| v_d \| = d^{\delta}$ with $\delta \in (1/4, 1/2)$, which correspond to the strong inconsistent case of Theorem 3.1 since $\| v_d \| / d^{1/2} \to 0$. That is, in some cases, the classifier $W(X_0)$ can have the consistency property of the error rates even when the normal vector is strong inconsistent with the optimal direction. This shows the good properties of the bias-corrected classifiers.

5 Simulation study

In this section we present a simulation study to illustrate numerically the theoretical results presented previously.

In Bolivar-Cime and Marron (2013) it is presented a simulation study considering Gaussian data with identity covariance matrix. In the simulations of the present paper we take more general multivariate data with an asymptotic geometric representation, to illustrate the asymptotic behavior of the four considered binary discrimination methods as the dimension tends to infinity. To compare our results with those of Bolivar-Cime and Marron (2013), we take the same mean vectors v_d considered in that paper. We take $v_d = (d^{\delta}, 0, \ldots, 0)^{\top}$ with $\delta = 0.2$, $\delta = 0.5$ and $\delta = 0.8$, which correspond to the cases $c = 0$, $c = 1$ and $c = \infty$ of Theorem 3.1, respectively. We also consider $v_d = \beta \mathbf{1}_d$ with $\beta = 0.5$, $\beta = 1$ and $\beta = 10$, which correspond to the cases $c = 0.5$, $c = 1$ and $c = 10$, respectively. We consider sample sizes $m = n = 20$, thus $\gamma = 1/m + 1/n = 1/10$. We take the dimensions $d = 10, 30, 100, 300, 1000, 2000$ to consider the non-HDLSS and HDLSS settings. The number of the training data sets generated for each value of d is $M = 500$.

In Ahn and Marron (2010) the MDP method is defined when $d \geq N - 1$, where $N = m + n$. They also mentioned that a formula for the normal vector of MDP is equivalent to Fisher's discriminant vector when $d \leq N - 2$, which does not have the piling property. We can view MDP as the HDLSS version of Fisher's linear discriminant method, with zero within-class scatter and maximized between-classes scatter. Hence, in the simulation study presented here, we take the MDP normal vector as the Fisher's discriminant vector when $d \leq N - 2$.

Suppose d is even. Let $Z = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(d)})^{\top}$ be a random vector where $Z^{(i)}$, for $i = 1, 2, \ldots, d/2$, are independent and identically distributed random variables from the univariate standard normal distribution, and where $Z^{(j)} = Z^{(i)2} + Z^{(i)} - 1$, for $j = d/2 + i$ with $i = 1, 2, \ldots, d/2$. Note that Z has mean zero and it can be seen that its covariance matrix is given by

$$\Sigma_d = \begin{bmatrix} \mathrm{I}_{d/2} & \mathrm{I}_{d/2} \\ \mathrm{I}_{d/2} & 3\mathrm{I}_{d/2} \end{bmatrix}. \tag{19}$$

Let $Z_1, Z_2, \ldots, Z_N$ be independent and identically distributed random vectors with the same distribution as Z. In Section 6.4 it is shown that these data have an asymptotic geometric representation, in particular the data satisfy condition (i) of Theorem 3.1 with $\sigma^2 = 2$.

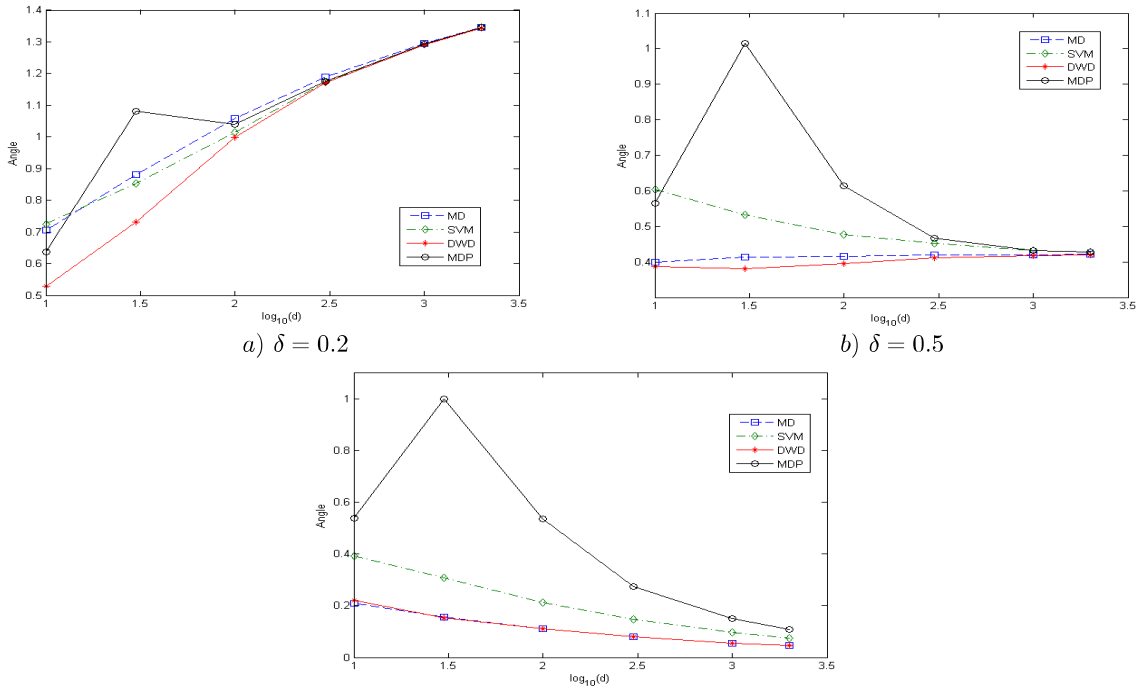
Now we will see that the data satisfy condition (ii) of Theorem 3.1. It is clear that the second moments of the entries of the vector Z_1 are uniformly bounded, since the distribution of the entries is only of two types (the standard normal distribution or the distribution of $Z^{(i)2} + Z^{(i)} - 1$, where $Z^{(i)}$ has standard normal distribution), and these distributions have finite second moments which are independent of the value of d. Observe that if $v_d = (d^\delta, 0, \ldots, 0)^\top$ with $\delta > 0$, by (II) of Remark 3.2 we have condition (ii) of Theorem 3.1. On the other hand, if $v_d = \beta \mathbf{1}_d$ with $\beta > 0$, by (IVb) of Remark 3.2 we also have the condition (ii) of Theorem 3.1, since Σ_d has a fixed number of nonzero upper diagonals.

5.1 Behavior of the normal vectors of the four methods

The means of the angles between v_d and the normal vectors of the separating hyperplanes of the four considered methods are computed for each value of d, in all the considered settings. In Figure 2 we show the means of the angles between v_d and the normal vectors varying the dimension d, for the case when $v_d = (d^\delta, 0, \ldots, 0)^\top$ with $\delta = 0.2$, $\delta = 0.5$ and $\delta = 0.8$. We observe that when $\delta = 0.2$ the means of the angles between the optimal vector v_d and the normal vectors of the separating hyperplanes of the four methods tend to approximate $\pi/2 = 1.5707$ as the dimension increases. When $\delta = 0.8$ the means of the angles tend to zero as the dimension increases. In this case, when $\delta = 0.5$ the means of the angles tend to $\arccos(c/(\gamma\sigma^2 + c^2)^{1/2}) = 0.4205$ as the dimension increases, where $c = 1$ ($\gamma = 1/10$ and $\sigma^2 = 2$). These results are according to Theorem 3.1.

In Figure 3 we show the means of the angles between v_d and the normal vectors of the four considered methods varying the dimension d, when $v_d = \beta \mathbf{1}_d$ with $\beta = 0.5,\ 1,\ 10$. It is observed that when β is equal to 0.5, 1 and 10 the means of the angles approximate 0.7297, 0.4205 and 0.0447 as the dimension increases, respectively, which are the values of $\arccos(c/(\gamma\sigma^2 + c^2)^{1/2})$ with c equal to 0.5, 1 and 10, respectively. Again, these results are according to Theorem 3.1.

For these data we observe that although the four methods have the same asymptotic behavior as the dimension tends to infinity, in terms of angles between the normal vectors and the optimal direction, the two best methods are MD and DWD, being MD the method with the best behavior in most of the cases. Note that in the case where all the methods are consistent, MD is the method that converge faster to the optimal direction, this is because the asymptotic optimal direction for the normal vector of the separating hyperplane is the difference between the two class means. In the HDLSS situation DWD some times coincide



a) $\delta = 0.2$ *b*) $\delta = 0.5$

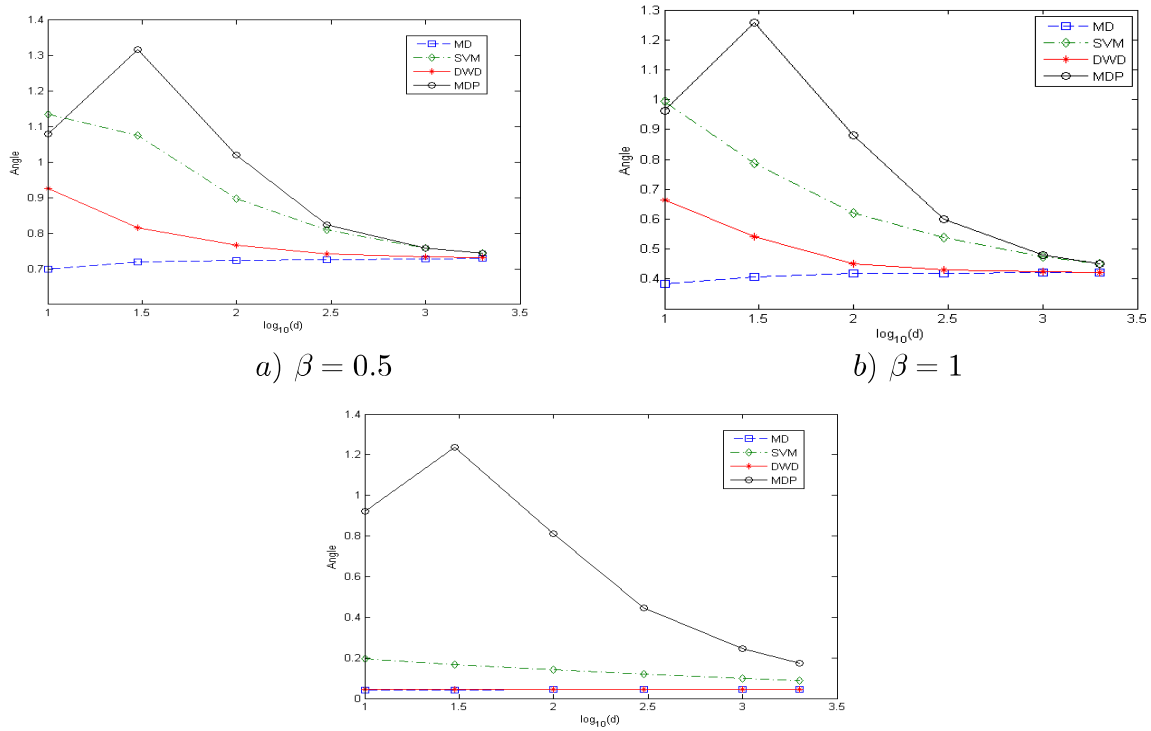
c) $\delta = 0.8$

Figure 2: Means of the angles between v_d and the normal vectors of the separating hyperplanes of the four methods, considering $v_d = (d^{\delta}, 0, \ldots, 0)^{\top}$ with $\delta = 0.2,\ 0.5,\ 0.8$.

with MD. The third best method is SVM and the worse method is MDP in almost all the considered cases. It is also observed that MDP has its worse behavior when $d \approx N = 40$, which has been previously noted in the simulations of Bolivar-Cime and Marron (2013), and in the simulations of Ahn and Marron (2010) for Gaussian data in terms of means of misclassification rates.

5.2 Behavior of the error rates of the four methods

In order to compare the error rates of the four methods we take $v_d = \beta \mathbf{1}_d$ with $\beta = 0.5,\ 1,\ 10$. Classification error rates were computed taking 100 new data points from each of the two classes. In Figure 4 we show the means of the error rates of the four considered methods for the cases $\beta = 0.5$ and $\beta = 1$, which correspond to the cases $c = 0.5$ and $c = 1$, respectively. In this figure we observe the convergence to zero of the means of the error rates of the four methods as $d \to \infty$, even when in Figure 3 we observed that the means of the angles of the normal vectors of the separating hyperplanes and the optimal direction do not converge to zero as $d \to \infty$. For the case $\beta = 10$, corresponding to the case $c = 10$, the means of the error rates of the four methods and for all the considered values of d were practically zero, therefore we do not include the graphs of the error rates for this case. In Figure 4 we observe that generally the error rates of the methods



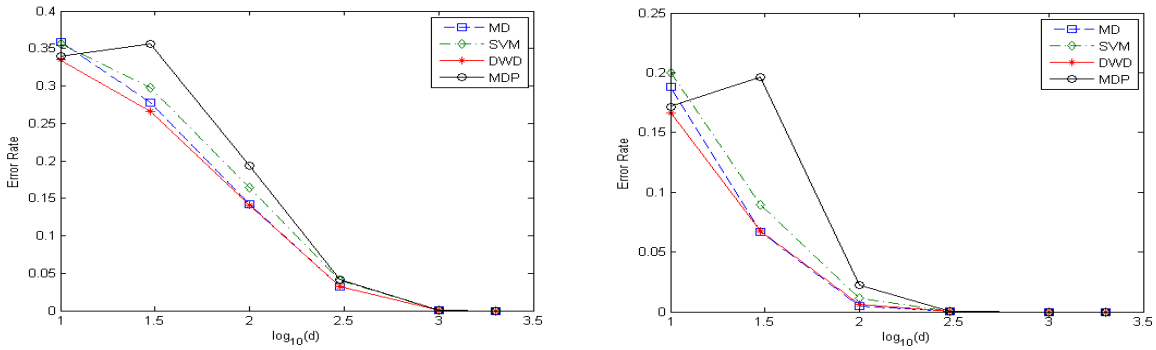
a) $\beta = 0.5$ *b)* $\beta = 1$

c) $\beta = 10$

Figure 3: Means of the angles between v_d and the normal vectors of the separating hyperplanes of the four methods, considering $v_d = \beta \mathbf{1}_d$ with $\beta = 0.5,\ 1,\ 10$.

DWD and MD are the smallest, and in the HDLSS situation these error rates practically coincide. The third best method in terms of error rates is SVM, and the worse method is MDP. Note that similar conclusions were obtained in Section 5.1 in terms of angles between the normal vectors of the methods and the optimal direction, however DWD sometimes have smaller error rates than MD, and MD generally have smaller angles than DWD. We also observe that as c increases the convergence of the error rates to zero is faster.

In Hall et al. (2005) it is studied by simulations the behavior of the error rates of the methods MD, SVM and DWD when n and m are unequal. They observed in their simulations that when d tends to infinity the error rates of MD and SVM practically coincide, and DWD es substantially worse than these methods, considering several values of $c > 0$. We did similar simulations but now including the MDP method, not showing here to save space, and we observed that when d tends to infinity the error rates of MDP practically coincide with error rates of MD and SVM. This was expected, since for $0 < c < \infty$ the hyperplanes of these three methods coincide as d tends to infinity due to the asymptotic geometric representation of the data, and by the results of Section 4 these three methods behave better than DWD in terms of error rates.



a) $\beta = 0.5$ *b)* $\beta = 1$

Figure 4: Means of the error rates of the four methods, considering $v_d = \beta \mathbf{1}_d$ with $\beta = 0.5,\ 1$.

6 Technical details

The main ideas for the proofs of our results are similar to that of Bolivar-Cime and Marron (2013), however we use the Tchebysheff and the Cauchy-Schwartz inequalities. First we present some consequences of the hypothesis of Theorem 3.1 that will be very useful along this section.

Let $\langle x, y\rangle = \sum_{k=1}^{d} x^{(k)} y^{(k)}$ be the inner product of the vectors $x, y \in \mathbb{R}^d$. By condition (i) we have

$$\frac{\langle Z_i, Z_j\rangle}{d} = \frac{1}{2}\left(\frac{\parallel Z_i \parallel^2}{d} + \frac{\parallel Z_j \parallel^2}{d} - \frac{\parallel Z_i - Z_j \parallel^2}{d}\right) \xrightarrow{P} 0, \qquad \text{for } i \neq j, \tag{20}$$

as $d \to \infty$. Furthermore, for any $i = 1, 2, \ldots, N$, by condition (ii) and the Tchebysheff inequality, for all $\tau > 0$

$$P\left(\frac{|\langle Z_i, v_d\rangle|}{d^{1/2} \parallel v_d \parallel} > \tau\right) \leq \frac{E(\langle Z_i, v_d\rangle^2)}{\tau^2 d \parallel v_d \parallel^2} = \frac{\sum_{k,l=1}^{d} E(Z_i^{(k)} Z_i^{(l)}) v_d^{(k)} v_d^{(l)}}{\tau^2 d \parallel v_d \parallel^2} = \frac{1}{\tau^2} \frac{\sum_{k,l=1}^{d} \sigma_{k,l} v_d^{(k)} v_d^{(l)}}{d \parallel v_d \parallel^2} \longrightarrow 0$$

as $d \to \infty$. Thus for all $i = 1, 2, \ldots, N$

$$\frac{\langle Z_i, v_d\rangle}{d^{1/2} \parallel v_d \parallel} \xrightarrow{P} 0 \qquad \text{as } d \to \infty. \tag{21}$$

Note that if $0 \leq c < \infty$ then (21) implies

$$\frac{\langle Z_i, v_d\rangle}{d} = \frac{\langle Z_i, v_d\rangle}{d^{1/2} \parallel v_d \parallel} \frac{\parallel v_d \parallel}{d^{1/2}} \xrightarrow{P} 0 * c = 0 \qquad \text{as } d \to \infty. \tag{22}$$

6.1 Proof of Lemma 3.1

Observe the following

$$\frac{\| X_i - Y_j \|^2}{d} = \frac{\| Z_i + v_d - Y_j \|^2}{d}$$
$$= \frac{\| Z_i \|^2}{d} + \frac{\| Y_j \|^2}{d} + \frac{\| v_d \|^2}{d} - 2\frac{\langle Z_i, Y_j \rangle}{d} + 2\frac{\langle Z_i, v_d \rangle}{d} - 2\frac{\langle Y_j, v_d \rangle}{d}.$$

Therefore, by (13), (20) and (22) it follows that

$$\frac{\| X_i - Y_j \|^2}{d} \xrightarrow{P} 2\sigma^2 + c^2 \qquad \text{as } d \to \infty. \tag{23}$$

We also have by (13) that for $i \neq j$

$$\frac{\| X_i - X_j \|^2}{d} \xrightarrow{P} 2\sigma^2, \qquad \frac{\| Y_i - Y_j \|^2}{d} \xrightarrow{P} 2\sigma^2 \qquad \text{as } d \to \infty. \tag{24}$$

Thus (23) and (24) imply that the vectors $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ tend to be the vertices of an N-polyhedron as $d \to \infty$. The rest of the proof is based on similar arguments as in the proof of Lemma 3.1 of Bolivar-Cime and Marron (2013), which are presented below.

The asymptotic N-polyhedron has m of its vertices arranged as those of an m-simplex and the other n vertices arranged in an n-simplex. After rescaling by $d^{-1/2}$, when d tends to infinity the data in C_+ and C_- tend to be the vertices of an m-simplex and an n-simplex, respectively, with edges of length $2^{1/2}\sigma$. Let $X_1^*, \ldots, X_m^*$ be the vertices of the m-simplex and let $Y_1^*, \ldots, Y_n^*$ be the vertices of the n-simplex. Let $\widetilde{v} = \mathbf{X}\alpha_+ - \mathbf{Y}\alpha_-$, with $\alpha \geq \mathbf{0}$ and $\mathbf{1}_m^\top \alpha_+ = \mathbf{1}_n^\top \alpha_- = 1$, be proportional to the normal vector of the MD, SVM or DWD hyperplane. For the two classes of the N-polyhedron, it is shown in the proof of Lemma 3.1 of Bolivar-Cime and Marron (2013) that this α is given by

$$\widehat{\alpha}_{i+} = \frac{1}{m}, \quad \widehat{\alpha}_{j-} = \frac{1}{n}, \quad i = 1, 2, \ldots, m, \ j = 1, 2, \ldots, n,$$

for these three methods. Therefore we have (17).

6.2 Proof of Theorem 3.1

6.2.1 When v is the normal vector of the MD, SVM or DWD hyperplane

CASE 1: $c = \infty$. Let $\widetilde{v} = \sum_{i=1}^{m} \alpha_{i+} X_i - \sum_{i=1}^{n} \alpha_{i-} Y_i$ be proportional to the vector v, with $\alpha > \mathbf{0}$ and $\mathbf{1}_m^\top \alpha_+ = \mathbf{1}_n^\top \alpha_- = 1$, as in Lemma 3.1. We have

$$\langle \widetilde{v}, v_d \rangle = \sum_{i=1}^{m} \alpha_{i+} \langle Z_i, v_d \rangle - \sum_{i=1}^{n} \alpha_{i-} \langle Y_i, v_d \rangle + \| v_d \|^2 . \tag{25}$$

Note that by the Cauchy-Schwartz inequality $| \langle Z_i, v_d \rangle | \leq \| Z_i \| \| v_d \|$, therefore

$$\frac{| \langle Z_i, v_d \rangle |}{\| v_d \|^2} \leq \frac{\| Z_i \|}{d^{1/2}} \frac{d^{1/2}}{\| v_d \|} \xrightarrow{P} \sigma * 0 = 0 \tag{26}$$

as $d \to 0$, for $i = 1, 2, \ldots, N$. Since $0 \leq \alpha_{i+}, \alpha_{i-} \leq 1$, it follows that

$$\frac{\sum_{i=1}^{m} \alpha_{i+} \langle Z_i, v_d \rangle}{\| v_d \|^2} \xrightarrow{P} 0, \quad \frac{\sum_{i=1}^{n} \alpha_{i-} \langle Y_i, v_d \rangle}{\| v_d \|^2} \xrightarrow{P} 0 \qquad \text{as } d \to \infty. \tag{27}$$

Thus, by (25) we have

$$\frac{\langle \widetilde{v}, v_d \rangle}{\| v_d \|^2} \xrightarrow{P} 1 \qquad \text{as } d \to \infty. \tag{28}$$

We also have

$$\| \widetilde{v} \|^2 = \sum_{i=1}^{d} \left(\sum_{i=1}^{m} \alpha_{i+} Z_i^{(k)} - \sum_{i=1}^{n} \alpha_{i-} Y_i^{(k)} \right)^2 + 2 \left(\sum_{i=1}^{m} \alpha_{i+} \langle Z_i, v_d \rangle - \sum_{i=1}^{n} \alpha_{i-} \langle Y_i, v_d \rangle \right) + \| v_d \|^2 \tag{29}$$

The first term of the last equation is equal to

$$\begin{aligned} &\sum_{i=1}^{m} \alpha_{i+}^2 \| Z_i \|^2 + 2 \sum_{i<j} \alpha_{i+} \alpha_{j+} \langle Z_i, Z_j \rangle + \sum_{i=1}^{n} \alpha_{i-}^2 \| Y_i \|^2 + 2 \sum_{i<j} \alpha_{i-} \alpha_{j-} \langle Y_i, Y_j \rangle \\ &\quad - 2 \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{i+} \alpha_{j-} \langle Z_i, Y_j \rangle . \end{aligned} \tag{30}$$

Using that $0 \leq \alpha_{i+} \leq 1$ and the Cauchy-Schwartz inequality we have

$$\begin{aligned} &\frac{\alpha_{i+}^2 \| Z_i \|^2}{\| v_d \|^2} \leq \frac{\| Z_i \|^2}{d} \frac{d}{\| v_d \|^2} \xrightarrow{P} \sigma^2 * 0 = 0 \qquad \forall i, \\ &\frac{|\alpha_{i+} \alpha_{j+} \langle Z_i, Z_j \rangle|}{\| v_d \|^2} \leq \frac{\| Z_i \|}{d^{1/2}} \frac{\| Z_j \|}{d^{1/2}} \frac{d}{\| v_d \|^2} \xrightarrow{P} \sigma^2 * 0 = 0 \qquad \text{for } i \neq j, \end{aligned}$$

as $d \to \infty$. Thus, if we divide the first two terms of (30) by $\| v_d \|^2$ they converge to zero in probability as $d \to \infty$. Analogously, dividing the rest of the terms of (30) by $\| v_d \|^2$ they converge to zero in probability as $d \to \infty$. Note that if we divide the second term in the right-hand side of (29) by $\| v_d \|^2$, due to (27) this term also converges to zero in probability as $d \to \infty$. Thus

$$\frac{\| \widetilde{v} \|^2}{\| v_d \|^2} \xrightarrow{P} 1 \qquad \text{as } d \to \infty. \tag{31}$$

By (28) and (31) we have

$$\frac{\langle \widetilde{v}, v_d \rangle}{\| \widetilde{v} \| \| v_d \|} = \frac{\langle \widetilde{v}, v_d \rangle / \| v_d \|^2}{\| \widetilde{v} \| / \| v_d \|} \xrightarrow{P} 1 \tag{32}$$

as $d \to \infty$. Therefore

$$\text{Angle}(\widetilde{v}, v_d) = \arccos\left(\frac{\langle \widetilde{v}, v_d \rangle}{\| \widetilde{v} \| \| v_d \|} \right) \xrightarrow{P} 0 \qquad \text{as } d \to \infty.$$

Note that for this case the condition (ii) is not necessary.

Case 2: $0 \leq c < \infty$. Let $\widetilde{v}$ be as before. By Lemma 3.1 and (13) it follows that

$$\sum_{i=1}^{m} \alpha_{i+}^2 \frac{\| Z_i \|^2}{d} \xrightarrow{P} \sum_{i=1}^{m} \frac{1}{m^2} \sigma^2 = \frac{\sigma^2}{m}, \qquad \sum_{i=1}^{m} \alpha_{i-}^2 \frac{\| Y_i \|^2}{d} \xrightarrow{P} \sum_{i=1}^{n} \frac{1}{n^2} \sigma^2 = \frac{\sigma^2}{n}, \tag{33}$$

as $d \to \infty$. We have that $\| \widetilde{v} \|^2$ is given by (29) and the first term of (29) is given by (30). Now, by Lemma 3.1, (20) and (33) we have that (30) divided by d converges in probability to $\gamma\sigma^2$, with $\gamma = \frac{1}{m} + \frac{1}{n}$, as $d \to \infty$. By Lemma 3.1 and (22) the second term of (29) divided by d converges in probability to zero as $d \to \infty$. Thus, since $\| v_d \|^2 / d \to c^2$ we have that

$$\frac{\| \widetilde{v} \|^2}{d} \xrightarrow{P} \gamma\sigma^2 + c^2 \qquad \text{as } d \to \infty. \tag{34}$$

Dividing both sides of (25) by $d^{1/2} \| v_d \|$, Lemma 3.1 and (21) imply

$$\frac{\langle \widetilde{v}, v_d \rangle}{d^{1/2} \| v_d \|} \xrightarrow{P} c \qquad \text{as } d \to \infty. \tag{35}$$

Now, by (34) and (35) we have

$$\frac{\langle \widetilde{v}, v_d \rangle}{\| \widetilde{v} \| \| v_d \|} = \frac{\langle \widetilde{v}, v_d \rangle / (d^{1/2} \| v_d \|)}{\| \widetilde{v} \| / d^{1/2}} \xrightarrow{P} \frac{c}{(\gamma\sigma^2 + c^2)^{1/2}} \tag{36}$$

as $d \to \infty$. Therefore

$$\text{Angle}(\widetilde{v}, v_d) = \arccos\left(\frac{\langle \widetilde{v}, v_d \rangle}{\| \widetilde{v} \| \| v_d \|}\right) \xrightarrow{P} \arccos\left(\frac{c}{(\gamma\sigma^2 + c^2)^{1/2}}\right) \tag{37}$$

as $d \to \infty$. Note that if $c = 0$ then $\arccos(c/(\gamma\sigma^2 + c^2)^{1/2}) = \pi/2$.

6.2.2 When v is the normal vector of the MDP hyperplane

Let $\overline{X}$ and $\overline{Y}$ be the mean vectors of the classes C_+ and C_-, respectively. Let $u = \overline{X} - \overline{Y}$ be the MD normal vector. First, we will show that for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ we have

$$\text{Angle}(X_i - \overline{X}, u) \xrightarrow{P} \frac{\pi}{2}, \qquad \text{Angle}(Y_j - \overline{Y}, u) \xrightarrow{P} \frac{\pi}{2} \qquad \text{as } d \to \infty. \tag{38}$$

Observe that

$$\begin{aligned} \| X_i - \overline{X} \|^2 &= \| (1 - 1/m)Z_i - (1/m) \sum_{j \neq i} Z_j \|^2 \\ &= \left(1 - \frac{1}{m}\right)^2 \| Z_i \|^2 - 2\left(1 - \frac{1}{m}\right)\left(\frac{1}{m}\right) \sum_{j \neq i} \langle Z_i, Z_j \rangle + \frac{1}{m^2} \sum_{k=1}^{d} (\sum_{j \neq i} Z_j^{(k)})^2. \end{aligned}$$

By (13) the first term of the last expression divided by d converges in probability to $(1 - 1/m)^2\sigma^2$ as $d \to \infty$. By (20) the second term divided by d converges in probability to zero as $d \to \infty$. Observe that the third term is equal to

$$\frac{1}{m^2}\left(\sum_{j \neq i} \| Z_i \|^2 + 2 \sum_{r,s \neq i,\ r<s} \langle Z_r, Z_s \rangle\right).$$

Therefore, (13) and (20) imply that the third term divided by d converges in probability to $(m-1)\sigma^2/m^2$ as $d \to \infty$. Thus we have that

$$\frac{\| X_i - \overline{X} \|^2}{d} \xrightarrow{P} \left(1 - \frac{1}{m}\right)^2 \sigma^2 + \frac{m-1}{m^2}\sigma^2 = \frac{m-1}{m}\sigma^2 \tag{39}$$

as $d \to \infty$. If $\overline{Z}$ is the mean vector of $Z_1, Z_2, \ldots, Z_m$, then from (13) and (20) it follows that

$$\frac{\| \overline{Z} \|^2}{d} \xrightarrow{P} \frac{\sigma^2}{m} \qquad \text{as } d \to \infty. \tag{40}$$

Observe that

$$\begin{aligned}&\left\langle X_i - \overline{X}, u\right\rangle = \left\langle Z_i - \overline{Z}, \overline{Z} + v_d - \overline{Y}\right\rangle \\ &= \frac{1}{m} \parallel Z_i \parallel^2 + \frac{1}{m}\sum_{j\neq i} \langle Z_i, Z_j\rangle + \langle Z_i, v_d\rangle - \left\langle Z_i, \overline{Y}\right\rangle - \parallel \overline{Z} \parallel^2 - \left\langle \overline{Z}, v_d\right\rangle + \left\langle \overline{Z}, \overline{Y}\right\rangle\end{aligned}$$

By the last equation, if $c = \infty$, from (13), (20), (21) and (40) it follows that

$$\frac{\left\langle X_i - \overline{X}, u\right\rangle}{d^{1/2} \parallel v_d \parallel} \xrightarrow{P} 0 \qquad \text{as } d \to \infty. \tag{41}$$

Therefore, by (39), (41) and since $\parallel u \parallel / \parallel v_d \parallel \xrightarrow{P} 1$ as $d \to \infty$ due to (31), we have that

$$\cos[\text{Angle}(X_i - \overline{X}, u)] = \frac{\left\langle X_i - \overline{X}, u\right\rangle}{\parallel X_i - \overline{X} \parallel \parallel u \parallel} = \frac{\left\langle X_i - \overline{X}, u\right\rangle / (d^{1/2} \parallel v_d \parallel)}{(\parallel X_i - \overline{X} \parallel / d^{1/2})(\parallel u \parallel / \parallel v_d \parallel)} \xrightarrow{P} 0 \tag{42}$$

as $d \to \infty$. Similarly, $\cos[\text{Angle}(Y_j - \overline{Y}, u)] \xrightarrow{P} 0$ as $d \to \infty$. Thus we have (38). Analogously, but now using (22) and (34), it can be shown that (38) is also true for $0 \leq c < \infty$. The rest of the proof is based on similar arguments as in the proof of Theorem 3.1 of Bolivar-Cime and Marron (2013), which are presented below.

Define C as the matrix whose columns are the vectors $X_i - \overline{X}$, $Y_j - \overline{Y}$, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. By Ahn and Marron (2010), the normal vector of the MDP method is given by $v = Qu / \parallel Qu \parallel$ where Q is the symmetric projection matrix on the orthogonal complement of the column space of C. By (38), u tends to be in the orthogonal complement of the column space of C. Therefore when d is large Qu can be approximated by u, and v can be approximated by $u / \parallel u \parallel$. Thus $\cos(\text{Angle}(v, v_d)) = \langle v, v_d\rangle / (\parallel v \parallel \parallel v_d \parallel)$ can be approximated by

$$\frac{\langle u, v_d\rangle}{\parallel u \parallel \parallel v_d \parallel}. \tag{43}$$

Hence, as it was shown in Section 6.2.1, (43) converges to 1 in probability if $c = \infty$, and it converges to $c/(\gamma\sigma^2 + c^2)^{1/2}$ if $0 \leq c < \infty$.

6.3 Assumptions that imply condition (ii) of Theorem 3.1

Now it is shown that each condition (I), (II), (III) and (IV) implies the condition (ii) of Theorem 3.1.

(I) In this case

$$\frac{\left|\sum_{k,r=1}^{d}\sigma_{k,r}v_d^{(k)}v_d^{(r)}\right|}{d\parallel v_d\parallel^2}=\frac{\sum_{k=1}^{d}\sigma_{k,k}v_d^{(k)2}}{d\parallel v_d\parallel^2}\leq\frac{M}{d}\longrightarrow 0$$

as $d\to\infty$, where M is a bound of the entries of Σ_d.

(II) We have that

$$\begin{aligned}\frac{\left|\sum_{k,r=1}^{d}\sigma_{k,r}v_d^{(k)}v_d^{(r)}\right|}{d\parallel v_d\parallel^2}&\leq\frac{\sum_{k=1}^{d}\sigma_{k,k}v_d^{(k)2}}{d\parallel v_d\parallel^2}+2\sum_{k<r}\frac{|\sigma_{k,r}||v_d^{(k)}||v_d^{(r)}|}{d\parallel v_d\parallel^2}\\&\leq\frac{RM}{d}+\frac{(R-1)RM}{d}\longrightarrow 0\end{aligned}$$

as $d\to\infty$, where R is the number of nonzero entries of v_d, and M is a bound of the second moments of the entries of Z_1. In the last inequality it is used that $|\sigma_{k,r}|\leq(\sigma_{k,k}\sigma_{r,r})^{1/2}\leq M$, and that $|v_d^{(k)}|\leq\parallel v_d\parallel$ for all k.

(III) Suppose that the entries of v_d are uniformly bounded by R, then

$$\frac{\left|\sum_{k,l=1}^{d}\sigma_{k,l}v_d^{(k)}v_d^{(r)}\right|}{d\parallel v_d\parallel^2}\leq R^2\frac{d}{\parallel v_d\parallel^2}\frac{\sum_{k,l=1}^{d}|\sigma_{k,l}|}{d^2}.\tag{44}$$

Since $d/\parallel v_d\parallel^2\to c^{-2}$ as $d\to\infty$, if

$$\frac{\sum_{k,l=1}^{d}|\sigma_{k,l}|}{d^2}\longrightarrow 0\qquad\text{as } d\to\infty\tag{45}$$

then the right-hand side of (44) tends to zero as $d\to\infty$. Now it is showed that (45) holds under conditions a), b) or c).

a) If M is a bound of the second moments of the entries of Z_1 then

$$\begin{aligned}\frac{\sum_{k,l=1}^{d}|\sigma_{k,l}|}{d^2}&=\frac{\sum_{k=1}^{d}\sigma_{k,k}}{d^2}+\frac{\sum_{0<|k-l|<r}|\sigma_{k,l}|}{d^2}+\frac{\sum_{|k-l|\geq r}|\sigma_{k,l}|}{d^2}\\&\leq\frac{M}{d}+\frac{(2d-r)(r-1)M}{d^2}+\frac{(d-r)(d-r+1)\rho(r)}{d^2},\end{aligned}\tag{46}$$

since $|\sigma_{k,r}|\leq(\sigma_{k,k}\sigma_{r,r})^{1/2}\leq M$. For any $\epsilon>0$, taking r large enough and later taking d sufficiently large, the right-hand side of (46) is less that ϵ. Therefore, since $\epsilon>0$ is arbitrary it follows (45).

b) If r is the number of nonzero upper diagonals of Σ_d, and M is a bound of the second moments of

the entries of Z_1, then

$$\frac{\sum_{k,l=1}^{d}|\sigma_{k,l}|}{d^2}\leq\sum_{k=1}^{d}\frac{\sigma_{k,k}}{d^2}+2\sum_{k<l}\frac{|\sigma_{k,l}|}{d^2}\leq\frac{M}{d}+\frac{2r(d-1)M}{d^2} \tag{47}$$

in the last inequality it is used that each upper diagonal has at most $(d-1)$ nonzero elements, and since there are r nonzero upper diagonals, there are at most $r(d-1)$ nonzero elements of Σ_d in the upper diagonals, and $|\sigma_{k,r}|\leq(\sigma_{k,k}\sigma_{r,r})^{1/2}\leq M$. Therefore, since the right-hand side of the last inequality of (47) tends to zero as $d\to\infty$ it follows (45).

c) Note the following

$$\frac{\sum_{k=1}^{d}\lambda_{k,d}^2}{d^2}=\frac{\sum_{k,l=1}^{d}[E(Z_i^{(k)}Z_i^{(l)})]^2}{d^2}=\frac{\sum_{k,l=1}^{d}\sigma_{k,l}^2}{d^2}. \tag{48}$$

Furthermore, if $\parallel x\parallel_p=(\sum_{k=1}^{n}|x^{(k)}|^p)^{1/p}$ is the p norm in $\mathbb{R}^n$, since $\parallel x\parallel_1\leq\sqrt{n}\parallel x\parallel_2$ we have that in $\mathbb{R}^{d^2}$

$$\sum_{k,l=1}^{d}|\sigma_{k,l}|\leq d\left(\sum_{k,l=1}^{d}\sigma_{k,l}^2\right)^{1/2},$$

then

$$\frac{\sum_{k,l=1}^{d}|\sigma_{k,l}|}{d^2}\leq\left(\frac{\sum_{k,l=1}^{d}\sigma_{k,l}^2}{d^2}\right)^{1/2}.$$

Therefore, by (16) and (48) the right-hand side of the last inequality tends to zero and we have (45).

(IV) In this case

$$\frac{\left|\sum_{k,r=1}^{d}\sigma_{k,r}v_d^{(k)}v_d^{(r)}\right|}{d\parallel v_d\parallel^2}=\frac{\left|\sum_{k,l=1}^{d}\sigma_{k,l}\right|}{d^2}\leq\frac{\sum_{k,l=1}^{d}|\sigma_{k,l}|}{d^2}.$$

Analogous to the case (III), by the last inequality if (45) holds it follows (14), and (45) follows from a), b) or c).

6.4 Asymptotic geometric representation of the data in the simulations

In this section we show that the multivariate data considered in the simulation studies have an asymptotic geometric structure as the dimension increases.

Since $E(Z_i^{(k)2})=1$ and $E[(Z_i^{(k)2}+Z_i^{(k)}-1)^2]=3$, for $i=1,2,\ldots,N$ and $k=1,2,\ldots,d/2$, by the Law

of Large Numbers (LLN) we have

$$\frac{\| Z_i \|^2}{d} = \frac{1}{2}\frac{\sum_{k=1}^{d/2} Z_i^{(k)2}}{d/2} + \frac{1}{2}\frac{\sum_{k=1}^{d/2}(Z_i^{(k)2} + Z_i^{(k)} - 1)^2}{d/2} \xrightarrow{P} \frac{1}{2} + \frac{3}{2} = 2$$

as $d \to \infty$. Analogously, since $E[(Z_i^{(k)} - Z_j^{(k)})^2] = 2$ and $E[(Z_i^{(k)2} + Z_i^{(k)} - Z_j^{(k)2} - Z_j^{(k)})^2] = 6$, for $i \neq j$ and $k = 1, 2, \ldots, d/2$, by the LLN we have

$$\begin{aligned}\frac{\| Z_i - Z_j \|^2}{d} &= \frac{1}{2}\frac{\sum_{k=1}^{d/2}(Z_i^{(k)} - Z_j^{(k)})^2}{d/2} + \frac{1}{2}\frac{\sum_{k=1}^{d/2}[Z_i^{(k)2} + Z_i^{(k)} - 1 - (Z_j^{(k)2} + Z_j^{(k)} - 1)]^2}{d/2} \\ &\xrightarrow{P} \frac{2}{2} + \frac{6}{2} = 4\end{aligned}$$

as $d \to \infty$. Thus the data have an asymptotic geometric representation and satisfy the condition (i) of Theorem 3.1 with $\sigma^2 = 2$.

7 Conclusions

The geometric representation of the HDLSS data allows to analyze the asymptotic behavior of some binary discrimination methods. In particular, under this geometric structure of the data and some conditions, we showed that Support Vector Machine, Distance Weighted Discrimination, Mean Difference and Maximal data Piling have the same asymptotic behavior, in terms of angles between the normal vectors of the separating hyperplanes and the optimal direction, as the dimension increases and the sample sizes are fixed. Our results generalize the results of Bolivar-Cime and Marron (2013), where it is showed that the four methods have the same asymptotic behavior in the Gaussian case where the classes have common diagonal covariance matrix.

Comparing the asymptotic behaviour of the angles between the normal vectors of the separating hyperplanes and the optimal direction with the asymptotic behavior of the error rates, we observed that in some cases the geometric representation of the data allows to have the consistency property of the error rates even when the normal vectors of the separating hyperplanes are inconsistent with the optimal direction. Due to the asymptotic geometric structure of the data, the classification with these four methods is easy when the distance between the two classes is large and it is more difficult when it is small, since as the distance between the two classes increases, the angles between the normal vectors and the optimal direction tend to zero and the error rates approach to zero faster when the dimension tends to infinity.

In the simulation study presented here, where the sample sizes of the two classes were the same, the conclusions in terms of angles between the normal vector and the optimal direction, and the conclusions in terms of error rates were similar: Although the four methods have the same asymptotic behavior as the

dimension tends to infinity, generally for large dimensions the two methods with the best behavior were MD and DWD, the third best method was SVM and the worse was MDP. The MD method had the best behavior in terms of angles of the normal vectors in most of the cases, this is because the asymptotic optimal direction for the normal vector of the separating hyperplane is the difference between the two class means. The results in terms of error rates for unequal sample sizes of the classes are totally different, since in this case as the dimension increases the methods MD, SVM and MDP practically coincide, and the DWD method is substantially worse than these methods. However, if the distance between the two classes is sufficiently large, the error rates of the four methods tend to zero as the dimension tends to infinity. As we have observed, the same asymptotic behavior of the error rates of the four methods is related with the same asymptotic behavior of the normal vectors of the four methods, given by the results presented here.

Acknowledgements

The authors are grateful to Aroldo Perez Perez and Victor Perez-Abreu for their help to improve an early version of this manuscript. They also thank the Editor Prof. N. Balakrishnan and the anonymous referees for their important comments and valuable suggestions, which helped greatly to improve this work. The authors thank the Universidad Juárez Autónoma de Tabasco for the support provided during the elaboration of this paper. This work was financed by PRODEP, Grant UJAT-PTC-178.

References

Ahn, J. and Marron, J. S. (2010). The Maximal Data Piling Direction for Discrimination. *Biometrika*, 97(1):254-259.

Ahn, J., Marron, J. S., Muller, K. M., and Chi, Y. (2007). The High-dimension, Low-sample-size Geometric Representation Holds Under Mild Conditions. *Biometrika*, 94(3):760-766.

Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Ann. Inst. Stat. Math.*, 66:983-1010.

Bolivar-Cime, A. and Marron, J. S. (2013). Comparison of binary discrimination methods for high dimension low sample size data. *J. Multivar. Anal.*, 115:108-121.

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

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, U.K.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric Representation of High Dimension, Low Sample Size Data. *J. R. Statist. Soc. B*, 67(3):427-444.

Jung, S. and Marron, J. S. (2009). PCA Consistency in High Dimension, Low Sample Size Context. *Ann. Statist.*, 37(6B):4104-4130.

Jung, S., Sen, A., and Marron, J. S. (2012). Boundary behavior in High Dimension, Low Sample Size asymptotics of PCA. *J. Multivar. Anal.*, 109:190-203.

Marron, J. S. (2015). Distance-weighted discrimination. *WIREs Comput. Stat.*, 7:109-114.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-Weighted Discrimination. *J. Am. Statist. Ass.*, 102(480):1267-1271.

Nakayama, Y., Yata, K., and Aoshima, M. (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *J. Stat. Plan. Inference*, in press. arXiv:1702.08019.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J., and Marron, J. S. (2010). Weighted Distance Weighted Discrimination and Its Asymptotic Properties. *J. Am. Statist. Ass.*, 105(489):401-414.

Qiao, X. and Zhang, L. (2015). Flexible high-dimensional classification machines and their asymptotic properties. *J. Mach. Learn. Res.*, 16:1547-1572.

Scholkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. The MIT press, Cambridge, Massachusetts.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.

Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivar. Anal.*, 105(1):193-215.