Journal of Multivariate Analysis 115 (2013) 108–121

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Comparison of binary discrimination methods for high dimension low sample size data

A. Bolivar-Cime [a,*], J.S. Marron [b]

[a] Department of Probability and Statistics, CIMAT, Jalisco S/N, Col. Valenciana, CP 36240, Guanajuato, Gto, Mexico

[b] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260, USA

ARTICLE INFO

Article history:
Received 10 June 2011
Available online 12 October 2012

AMS 2000 subject classifications:
primary 62H30
secondary 62E20

Keywords:
Asymptotic analysis
Binary discrimination
High dimensional data
Machine learning

ABSTRACT

A comparison of some binary discrimination methods is done in the high dimension low sample size context for Gaussian data with common diagonal covariance matrix. In particular we obtain results about the asymptotic behavior of the methods Support Vector Machine, Mean Difference (i.e. Centroid Rule), Distance Weighted Discrimination, Maximal Data Piling and Naive Bayes when the dimension d of the data sets tends to infinity and the sample sizes of the classes are fixed. It is concluded that, under appropriate conditions, the first four methods are asymptotically equivalent, but the Naive Bayes method can have a different asymptotic behavior when d tends to infinity.

1. Introduction

This work deals with *binary discrimination analysis* in the High-Dimension, Low Sample Size (HDLSS) framework for Gaussian data. We focus on the study of asymptotic behavior of the following methods for two-class discrimination: Support Vector Machine (SVM), Mean Difference (MD), Distance Weighted Discrimination (DWD), Maximal Data Piling (MDP) and Naive Bayes (NB). The HDLSS asymptotics of the first three methods have been previously studied in Hall et al. [7], where the probability of correct classification of a new data point is considered when the dimension d of the training data sets tends to infinity for fixed sample sizes of the classes. The present paper takes a different asymptotic viewpoint, based on the angle between the normal vectors of the separating hyperplane. We find conditions that characterize both consistency and strong inconsistency. Previous comparison of these methods has been done by simulations in Marron et al. [10], and Ahn and Marron [1]. A further contribution of the present paper is theoretical analysis of some empirical phenomena observed in [10,1], by specifically studying the asymptotic behavior of the normal vectors to the separating hyperplanes of these three methods, as the data dimension increases.

We give a description of the methods mentioned above in Section 2. In Section 3 we show that the first four methods have asymptotically the same first order behavior when the dimension d of the data sets tends to infinity, for fixed sample sizes of the classes. Specifically, we see that when the data sets are Gaussian with common diagonal covariance matrix and one set has mean zero and the other has mean v_d, then the normal vectors of the separating hyperplanes tend to be in the same direction as v_d when $\|v_d\| \gg d^{1/2}$, i.e. are *consistent*, and tend to be orthogonal to v_d when $\|v_d\| \ll d^{1/2}$, i.e. are *strongly inconsistent*. This paper also contains the HDLSS analysis of behavior in the interesting boundary case $\|v_d\| \approx d^{1/2}$. Moreover we observe that the NB method may have a different asymptotic behavior from the other four methods and may

* Corresponding author.
E-mail addresses: addy@cimat.mx (A. Bolivar-Cime), marron@email.unc.edu (J.S. Marron).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.10.001

be inconsistent in many situations where the other methods are consistent. We talk about some simulations that we have done to assess the theoretical results of this paper in Section 4. In Section 5 we give a discussion of our results. Finally, in Section 6 we give the proofs of the results presented in Section 3.

2. Binary discrimination methods

In this section we present the linear classification methods treated in this paper, which are based on separating hyperplanes. Suppose that we have the following training data set

$$(x_1, w_1), (x_2, w_2), \ldots, (x_N, w_N), \tag{1}$$

where $x_i \in \mathbb{R}^d$ and $w_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$. In particular, we have two classes of data, the classes C_+ and C_- corresponding to the vectors with $w_i = 1$ and $w_i = -1$ respectively. Let $X = [x_1, x_2, \ldots, x_N]$ be the $d \times N$ matrix of training data and $w = (w_1, w_2, \ldots, w_N)^\top$ be the vector of corresponding class labels. The following notation will be used:

- W is the $N \times N$ diagonal matrix with the elements of w on the diagonal,
- X_+ (X_-) is the sub-matrix of X corresponding to the class C_+ (C_-),
- m (n) is the cardinality of the class C_+ (C_-),
- $\mathbf{1}_k$ is the k-dimensional vector of ones.

We say that the training data set (1) is *linearly separable* if there exists a hyperplane for which all the data of the class C_+ are on one side of the hyperplane and all the data of the class C_- are on the other side. In this case a hyperplane with such property is called a *separating hyperplane* of the training data set.

Only the separable case is treated here, because in HDLSS situations the data from continuous probability densities are linearly separable almost surely (see [7]), and we consider multivariate Gaussian data in this paper.

2.1. Support Vector Machine

An introduction to the Support Vector Machine (SVM) method for binary discrimination analysis is given in this section. For more comprehensive and detailed studies see for example [4–6,8,15,16].

The SVM method was proposed by Vapnik in [15,16]. It is one of the most popular binary discrimination methods in the literature. In the linearly separable case, the SVM method consists of finding a separating hyperplane that maximizes the distances of the hyperplane to the nearest vector of each class.

Here we develop SVM from several view points, which will be needed in the following analysis. Suppose that there exist a vector v and a scalar b such that the following inequalities hold:

$$\begin{aligned} &v^\top x_i + b \geq 1, \quad \text{if } w_i = 1, \\ &v^\top x_i + b \leq -1, \quad \text{if } w_i = -1. \end{aligned} \tag{2}$$

In this case the hyperplane

$$v^\top x + b = 0$$

is a separating hyperplane of the training data set. Note that the inequalities in (2) can be written as

$$w_i(v^\top x_i + b) \geq 1, \quad i = 1, 2, \ldots, N. \tag{3}$$

The vectors x_i that satisfy the equality in (3) are called *support vectors*. That is, the support vectors are the training vectors that belong to one of the hyperplanes

$$v^\top x + b = -1 \quad \text{or} \quad v^\top x + b = 1. \tag{4}$$

The set of support vectors will be denoted by SV.

Let d_+ and d_- be the shortest distances from the separating hyperplane to the nearest vector in C_+ and C_-, respectively. Then the *margin* of the separating hyperplane is defined as $d_0 = d_+ + d_-$. Hence, the margin of the separating hyperplane is the distance between the hyperplanes given in (4) which is

$$d_0 = \frac{2}{\|v\|}.$$

In the separable case the *optimal separating hyperplane* or *SVM hyperplane*

$$v_0^\top x + b_0 = 0$$

is the unique separating hyperplane with a maximal margin. Thus the SVM hyperplane maximizes $2/\|v\|$ subject to the conditions (3). Equivalently, the SVM hyperplane solves the optimization problem

$$\begin{aligned} &\text{minimize } \frac{\|v\|^2}{2}, \\ &\text{subject to } w_i(v^\top x_i + b) \geq 1, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{5}$$

According to Burges [4], Cortes and Vapnik [5] and Izenman [8], the optimal vector is given by

$$v_0 = XW\widehat{\alpha}, \tag{6}$$

where $\widehat{\alpha}$ solves the optimization problem

$$\begin{aligned} &\text{maximize } \mathbf{1}_N^\top \alpha - \frac{1}{2}\|XW\alpha\|, \\ &\text{subject to } \alpha \geq \mathbf{0},\ \alpha^\top w = 0. \end{aligned} \tag{7}$$

Furthermore $\widehat{\alpha}_i \neq 0$ only if x_i is a support vector, hence by (6) v_0 is a linear function of the support vectors only. Since the support vectors satisfy the equality (3), the bias b_0 can be calculated as

$$b_0 = w_i - v_0^\top x_i, \tag{8}$$

for any $x_i \in SV$.

From [11] we have that the optimization problem (7) is equivalent to the following

$$\begin{aligned} &\text{maximize } \frac{2}{\|XW\alpha^*\|^2} \\ &\text{subject to } \mathbf{1}_m^\top \alpha_+^* = \mathbf{1}_n^\top \alpha_-^* = 1, \quad \alpha_+^*, \alpha_-^* \geq \mathbf{0}, \end{aligned} \tag{9}$$

where α_+^* and α_-^* are the sub-vectors of α^* corresponding to the classes C_+ and C_-. respectively. Note that $XW\alpha^* = X_+\alpha_+^* - X_-\alpha_-^*$; thus we are minimizing the distance between points in the convex hulls of the classes C_+ and C_-. Therefore if $\widehat{\alpha}^*$ solves the optimization problem (9) the normal vector of the SVM hyperplane can be taken as

$$v_0^* = XW\widehat{\alpha}^* = X_+\widehat{\alpha}_+^* - X_-\widehat{\alpha}_-^*, \tag{10}$$

which is the difference of a pair of closest points of the convex hulls and is proportional to the vector v_0 given by (6).

For the non-separable case see [4,5,8].

2.2. Distance Weighted Discrimination

In the HDLSS situation Marron et al. [10] observe that the projection of the data onto the normal vector of the SVM separating hyperplane produces substantial data piling (that is, many of these projections are the same), and they show that data piling may affect the *generalization performance* (how well new data from the same distributions can be discriminated). Therefore, they propose the Distance Weighted Discrimination (DWD) method, which avoids the data piling problem and improves generalizability. The idea of this method is to find a separating hyperplane that minimizes the sum of the reciprocals of the distances of the training data to the hyperplane. Thus, all the training data have a role in finding the hyperplane, but data close to the hyperplane have a bigger impact than data that are farther away. Here we describe briefly how this method works for the case when training data are linearly separable. The non-separable case can be found in [10,11].

Let $v \in \mathbb{R}^d$ be the normal vector of the separating hyperplane and $b \in \mathbb{R}$ its bias. Define the *residual* of the i-th data vector as

$$r_i = w_i(v^\top x_i + b). \tag{11}$$

The *DWD hyperplane*

$$v_1^\top x + b_1 = 0,$$

solves the optimization problem

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{N} \frac{1}{r_i} \\ &\text{subject to } \|v\| = 1, \quad r_i \geq 0,\ i = 1, 2, \ldots, N. \end{aligned} \tag{12}$$

As can be seen in [11], the optimal vector v_1 is given by

$$v_1 = \frac{XW\widehat{\beta}}{\|XW\widehat{\beta}\|} \tag{13}$$

where $\widehat{\beta}$ solves the optimization problem

$$\begin{aligned} &\text{maximize } 2\mathbf{1}_N^\top \sqrt{\beta} - \|XW\beta\|, \\ &\text{subject to } \beta \geq \mathbf{0}, \quad \beta^\top w = 0, \end{aligned} \tag{14}$$

with $\sqrt{\beta}$ denoting the vector whose components are the square roots of the components of β. Note that the optimization problem (14) is similar to that of (7) for the SVM method. On the other hand, from [11] the residuals are given by

$$r_i = \frac{1}{\sqrt{\widehat{\beta}_i}}, \quad i = 1, 2, \ldots, N. \tag{15}$$

Thus, from Eq. (11) the bias can be calculated as

$$b_1 = \frac{w_i}{\sqrt{\widehat{\beta}_i}} - v_1^\top x_i, \tag{16}$$

for any data vector x_i.

Similar to the case of the SVM method, [11] shows that the optimization problem (14) is equivalent to

$$\begin{aligned} &\text{maximize } \frac{(\mathbf{1}_m^\top \sqrt{\beta_+^*} + \mathbf{1}_n^\top \sqrt{\beta_-^*})^2}{\|X_+\beta_+^* - X_-\beta_-^*\|}, \\ &\text{subject to } \mathbf{1}_m^\top \beta_+^* = \mathbf{1}_n^\top \beta_-^* = 1, \quad \beta_+^*,\ \beta_-^* \geq \mathbf{0}. \end{aligned} \tag{17}$$

Hence we are trying to minimize the distance between points in the two convex hulls, but divided by the square of the sum of the square roots of the convex weights. Therefore if $\widehat{\beta}^*$ solves the optimization problem (17) the normal vector of the DWD hyperplane is proportional to

$$v_1^* = X_+\widehat{\beta}_+^* - X_-\widehat{\beta}_-^*. \tag{18}$$

Interesting generalizations of DWD and SVM have been proposed recently; see for example Qiao et al. [12] for weighted versions of DWD and SVM.

2.3. Mean Difference hyperplane

In the Mean Difference (MD) method, also called the nearest centroid method (see [13]), the separating hyperplane is the one that orthogonally bisects the segment joining the centroids or means of the two classes. That is, if the means of the classes C_+ and C_- are given by

$$\bar{x}_+ = \frac{1}{m}\sum_{x_i \in C_+} x_i \quad \text{and} \quad \bar{x}_- = \frac{1}{n}\sum_{x_i \in C_-} x_i, \tag{19}$$

respectively, then the *MD hyperplane* has normal vector

$$u = \bar{x}_+ - \bar{x}_- \tag{20}$$

and bisects the segment joining the means $\bar{x}_+$ and $\bar{x}_-$. Thus, as in the case of the SVM and DWD hyperplanes the normal vector of the MD hyperplane is the difference between two points on the convex hulls of the two classes.

2.4. Maximal Data Piling

The Maximal Data Piling (MDP) method for binary discrimination was proposed by Ahn and Marron [1]. This method was specially designed for the HDLSS context and we need to assume $d \geq N - 1$ and that the subspace generated by the data set has dimension $N - 1$. Under these assumptions there exist direction vectors onto which the projections of the training data are piled completely at two distinct points, one for each class. The normal vector of the MDP method is the direction vector for which the distance between these two points is maximal. On the other hand, in [1] it is shown that the MDP method is equivalent to the Fisher Linear Discrimination (FLD) in the non-HDLSS situation. However, Bickel and Levina [2] have demonstrated that FLD has very poor HDLSS properties, while in [1] it is shown that, although data piling may not be desirable, the MDP method can work very well and better than FLD in some settings in the HDLSS context.

In order to explain how to build the separating hyperplane for this method, let $u = \bar{x}_+ - \bar{x}_-$ be the MD normal vector given by (20). Define $C = [X_c^+, X_c^-]$, where X_c^+ and X_c^- are the centered versions of X_+ and X_- respectively, that is

$$X_c^+ = X_+ - \bar{x}_+\mathbf{1}_d^\top, \qquad X_c^- = X_- - \bar{x}_-\mathbf{1}_d^\top.$$

The symmetric projection matrix onto the orthogonal complement of the column space of C is given by $Q = I_d - CC^\dagger$, where $A^\dagger$ is the Moore–Penrose generalized inverse of A.

The *MDP hyperplane*

$$v_2^\top x + b_2 = 0$$

has the property that its unit normal vector v_2 is the direction for which the projections of the two class means have maximal distance, subject to the constraint that the projection of each training data onto the vector is the same as its class mean. In other words, v_2 solves the optimization problem

$$\begin{aligned} &\text{maximize } (v^\top u)^2, \\ &\text{subject to } C^\top v = 0, \qquad v^\top v = 1. \end{aligned} \tag{21}$$

It is seen in [1] that the normal vector is given by

$$v_2 = \frac{Qu}{\|Qu\|}. \tag{22}$$

This means that v_2 is orthogonal to the $N-2$ dimensional subspace generated by the columns of C. Furthermore, the formula (22) is equivalent to

$$v_2 = \frac{(X_c X_c^\top)^\dagger u}{\|(X_c X_c^\top)^\dagger u\|}, \tag{23}$$

where X_c is the centered version of the data matrix X. Therefore v_2 is also in the $N-1$ dimensional subspace generated by the globally centered data vectors. Finally, the bias b_2 can be calculated as

$$b_2 = -v_2^\top (m\bar{x}_+ + n\bar{x}_-)/N. \tag{24}$$

2.5. Naive Bayes

The Naive Bayes (NB) method assumes that the common covariance matrix Σ of the two classes is diagonal, i.e. the entries of the random vectors are uncorrelated. This method uses the Bayes Rule to obtain the normal vector of the separating hyperplane. Assuming that the classes have normal distributions $N_d(\mu_0, \Sigma)$ and $N_d(\mu_1, \Sigma)$, by [2] the normal vector of the *NB hyperplane* is given by

$$v_3 = D^{-1}u, \tag{25}$$

where u is the MD normal vector given by (20), and $D = \operatorname{diag}(\widehat{\Sigma})$ is the diagonal matrix whose entries are the diagonal elements of the *pooled covariance matrix*

$$\widehat{\Sigma} = \frac{1}{m+n-2}\left[\sum_{x_i \in C_+} (x_i - \bar{x}_+)^2 + \sum_{x_i \in C_-} (x_i - \bar{x}_-)^2\right]. \tag{26}$$

The bias of the NB hyperplane can be calculated as

$$b_3 = -v_3^\top \widehat{\mu}, \tag{27}$$

where $\widehat{\mu} = (\bar{x}_+ + \bar{x}_-)/2$.

3. Asymptotic results for the normal vectors

In this section we consider the $d \times N$ matrix $X = [x_1, x_2, \ldots, x_N]$ whose columns are a training data set for two classes. Suppose that the first m columns of X are the vectors of the class C_+ and the remaining $n = N - m$ columns are the vectors of the class C_-. Therefore, the matrices

$$\begin{aligned} X_+ &= [x_1, x_2, \ldots, x_m], \\ X_- &= [x_{m+1}, x_{m+2}, \ldots, x_{m+n}] \end{aligned} \tag{28}$$

are the sub-matrices of X corresponding to the classes C_+ and C_- respectively. We assume that the random vectors in C_+ and C_- are independent with d-multivariate normal distributions $N_d(v_d, \Sigma_d)$ and $N_d(0, \Sigma_d)$ respectively, where Σ_d is a diagonal matrix. Note that the difference between these classes is determined by the mean vector v_d. So the length of v_d, i.e. $\|v_d\|$, is crucial for classification performance.

As it was mentioned before, in the HDLSS setting the training data sets from continuous probability densities are linearly separable almost surely; see [7]. Thus, for $d > N$ we can assume that there exists a separating hyperplane for the classes C_+ and C_-.

Because the separating hyperplanes of the discrimination methods described in Section 2 are determined by their normal vectors, the behavior of classification is studied by the direction of these vectors. HDLSS asymptotic performance of all these methods will be related with the distance between the two class distributions, in particular by $\|v_d\|$. When $\|v_d\|$ is large classification becomes easy, and it is very challenging when $\|v_d\|$ is small. In view of the results of Hall et al. [7], who showed

that under certain conditions HDLSS data tend to lie at a distance $d^{1/2}$ from the mean when d is large, it is not surprising that $\|v_d\| \approx d^{1/2}$ is a critical boundary. This is confirmed by the next theorem which considers a common diagonal covariance matrix for the two classes of the training data set. When $\|v_d\| \gg d^{1/2}$, all the linear methods are *consistent* in the sense that the angles between the normal vectors of the hyperplanes and the optimal vector v_d converge to zero. When $\|v_d\| \ll d^{1/2}$, the normal vectors do not converge to the optimal direction. Furthermore they are *strongly inconsistent*, in the sense that they are asymptotically orthogonal. These results complement the results of classification from Hall et al. [7], which were formulated in terms of misclassification probabilities there, and normal vectors angles here. We further explore the boundary behavior as follows.

Theorem 3.1. *Suppose that the random vectors in C_+ and C_- are independent with d-multivariate normal distributions $N_d(v_d, \Sigma_d)$ and $N_d(\mathbf{0}, \Sigma_d)$ respectively, where*

$$\Sigma_d = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2) \tag{29}$$

and $\{\sigma_k\}_{k=1}^{\infty}$ is a bounded sequence of positive numbers such that

$$\sum_{k=1}^{d} \frac{\sigma_k^2}{d} \longrightarrow \sigma^2 \quad \text{as } d \to \infty, \tag{30}$$

for some $\sigma > 0$. Assume that $\|v_d\| d^{-1/2} \to c$ as $d \to \infty$, with $0 \le c \le \infty$. If v represents the normal vector of the MD, SVM, DWD or MDP hyperplane of the training data set, then

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

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

Observe that some matrices Σ_d satisfying the conditions of the last theorem are the positive multiples of the d-dimensional identity matrix, i.e. $\Sigma_d = \sigma^2 I_d$ for some $\sigma > 0$. Other examples are the diagonal matrices (29) for which the sequence of positive numbers $\{\sigma_k\}_{k=1}^{\infty}$ converges to a positive constant σ as $d \to \infty$, since the Cesaro means preserve convergent sequences and their limits.

In the simulation studies of Marron et al. [10] in terms of proportion of misclassification of new data points, it was observed that the MD, SVM and DWD methods have a similar behavior as dimension tends to infinity for fixed sample sizes and for Gaussian data with identity covariance matrix; see their Fig. 2(a). This is also observed in analogous simulations of Ahn and Marron [1], where MDP is also considered and has a similar asymptotic behavior than these three methods when dimension increases; see their Fig. 2 (first row). Now, Theorem 3.1 gives a theoretical explanation of this empirical result, since it claims that the normal vectors of the separating hyperplanes of these four methods have the same asymptotic behavior as the dimension tends to infinity. Furthermore, the asymptotic consistency and strong inconsistency of the methods are characterized by the asymptotic properties of the distance between the two classes, in particular by the asymptotic properties of $\|v_d\|$.

Although the normal vectors are asymptotically inconsistent when $0 \le c < \infty$, the next lemma gives an explicit asymptotic representation in this case, for the normal vectors of the three methods MD, SVM and DWD.

Lemma 3.1. *Under the assumptions of Theorem 3.1, let X_+ and X_- be as in (28) and assume $\|v_d\| d^{-1/2} \to c$, with $0 \le c < \infty$. If the vector $\widetilde{v} = X_+\alpha_+ - X_-\alpha_-$, with $\alpha \ge \mathbf{0}$ 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*

$$\alpha_{i,+} \xrightarrow{w} \frac{1}{m}, \qquad \alpha_{j,-} \xrightarrow{w} \frac{1}{n}, \tag{31}$$

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

We have not found such representation for MDP, but from the proof of Theorem 3.1, in Section 6, it follows that the unit length normal vector of the MDP hyperplane converges to the same direction as the normal vector of the other three methods as $d \to \infty$, since it can be approximated by the unit length normal vector of the MD hyperplane when d is large.

On the other hand, in the next proposition we show that the NB method can have a different asymptotic behavior from the other four methods. We see a case where the vector v_d satisfies $\|v_d\| d^{-1/2} \to c$ as $d \to \infty$, with $0 \le c \le \infty$, and the NB normal vector is inconsistent even if $c = \infty$. We also observe a case where the NB normal vector is consistent when $c = \infty$, and is strongly inconsistent when $c = 0$; i.e. the NB method has the same behavior as the other four methods for these values of c. However, when $c = 1$ the angle between v_d and the NB normal vector converges to a non-degenerate distribution as d tends to infinity. That is, the asymptotic behavior of this method not only depends on the distance between classes but also depends on the form of the mean vector v_d.

Proposition 3.1. *Assume the same as in Theorem 3.1. Furthermore, suppose that $R < \sigma_k < M\ \forall k = 1, 2, \ldots$ for some $R, M > 0$, and*

$$\sum_{k=1}^{d} \frac{1}{d\sigma_k^2} \longrightarrow \rho_1, \qquad \sum_{k=1}^{d} \frac{1}{d\sigma_k^4} \longrightarrow \rho_2 \tag{32}$$

as $d \to \infty$, for some $\rho_1, \rho_2 > 0$. Let v be the NB normal vector and assume $m + n > 6$.

(a) *Suppose $v_d = \beta \mathbf{1}_d$ where $\beta = \beta_d \to c$ as $d \to \infty$, with $0 \le c \le \infty$, then*

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} \arccos\left(\dfrac{\widetilde{\mu}}{(\widetilde{\sigma}^2 + \widetilde{\mu}^2)^{1/2}\rho_2^{1/2}}\right), & \text{if } c = \infty; \\ \arccos\left(\dfrac{c\widetilde{\mu}\rho_1}{(\gamma\rho_1 + c^2\rho_2)^{1/2}(\widetilde{\sigma}^2 + \widetilde{\mu}^2)^{1/2}}\right), & \text{if } 0 \le c < \infty; \end{cases} \tag{33}$$

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

$$\widetilde{\mu} = \frac{m+n-2}{m+n-4}, \qquad \widetilde{\sigma}^2 = \frac{2(m+n-2)^2}{(m+n-4)^2(m+n-6)}. \tag{34}$$

(b) *Suppose $v_d = (d^{\delta}, 0, \ldots, 0)^{\top}$, with $\delta > 0$. Then*

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} 0, & \text{if } \delta > 1/2; \\ \dfrac{\pi}{2}, & \text{if } \delta < 1/2; \\ \arccos([1 + \gamma(\widetilde{\sigma}^2 + \widetilde{\mu}^2)\rho_1\sigma_1^4\widetilde{s}^2]^{-1/2}), & \text{if } \delta = 1/2; \end{cases} \tag{35}$$

as $d \to \infty$, where $\widetilde{s}$ is a random variable with distribution $\chi^2_{(m+n-2)}/(m+n-2)$.

We observe that the NB normal vector is always inconsistent in case (a), being strongly inconsistent when $c = 0$. While in case (b) it has a similar behavior to the other four methods studied in this paper when $c = 0$ ($\delta < 1/2$) and $c = \infty$ ($\delta > 1/2$). However, the NB method has a very different asymptotic behavior from the other four methods when $c = 1$ ($\delta = 1/2$), since the angle between v_d and the NB normal vector converges to a non-degenerate distribution with support $[0, \pi/2]$.

Therefore, the consistency of the NB method is more difficult for case (a) than (b) of Proposition 3.1. This is because, in case (a), the mean vector must interact with all of the estimated marginal variances. This introduces a large amount of noise into the classification process, giving the poor performance shown in (33). On the other hand, in case (b), only one estimated marginal variance has substantial influence on the classification, therefore for the cases $\delta > 1/2$ and $\delta < 1/2$ its effect is asymptotically negligible. Thus the performance of the NB method is similar to the other classification methods in these cases.

Due to the asymptotic geometric properties of the data, the NB method has a strange behavior in the boundary case, $\|v_d\| \approx d^{1/2}$. In (a) and (b) of Proposition 3.1 we have seen that in the boundary case, $c = 1$ and $\delta = 1/2$ respectively, the NB angle converges to a constant in (a) and has a non-degenerate asymptotic distribution in (b), as dimension increases. This is because, in (a), the noise produced by the estimated marginal variances is averaged, while in (b), the mean vector only interacts with one marginal variance and its noise remains in the computation of the NB angle.

4. Simulation studies

In the supplementary material [3] of this paper we present some simulations where the five considered methods are compared in terms of the angles between the optimal direction v_d and the normal vectors of the separating hyperplanes; for dimensions $d = 10, 30, 100, 300, 1000, 2000$, and sample sizes of the two classes $m = n = 20$. We consider Gaussian data with identity covariance matrix. The mean v_d is taken as in (a) and (b) of Proposition 3.1 for some values of β and δ, respectively. These simulations consider the cases $c = 0, 0.5, 1, 10$ and $c = \infty$.

The results obtained in the simulations are according to the theoretical results presented in this paper. Specifically, in (a) and (b) of Fig. 1 it is observed the asymptotic strong inconsistency and consistency, respectively, of the five methods as dimension increases, given by Theorem 3.1 (cases $c = 0$ and $c = \infty$) and Proposition 3.1(b) (cases $\delta < 1/2$ and $\delta > 1/2$). In Fig. 2(a), we observe the inconsistency of the MD, SVM, DWD and MDP methods given by Theorem 3.1 (case $c = 1$). Furthermore, in Fig. 2(b), which corresponds to the case $\delta = 1/2$ of Proposition 3.1(b), it is shown that when dimension increases the box plots of the NB angles do not tend to be concentrated in a single value, therefore the NB angle has a non-degenerate asymptotic distribution. Finally, in (a) and (b) of Fig. 3 it is observed the asymptotic inconsistency of the methods given by Theorem 3.1 (cases $c = 0.5$ and $c = 10$) and Proposition 3.1(a) (cases $c = 0.5$ and $c = 10$).

In all the considered cases we see that the MD, SVM, DWD and MDP methods have the same asymptotic behavior. It is also observed that the MD method has the best behavior in all the cases and for all the values of d, in the sense that it has the

smallest mean of the angles between v_d and the normal vectors of the separating hyperplanes. This is not surprising because the MD is the likelihood ratio classifier in the Gaussian setting. Among the four methods that have the same asymptotic behavior, it is observed in the simulations that the second best method is DWD, the third is SVM, and the worst is MDP. Furthermore, we see that the NB method sometimes behaves asymptotically as the other four methods, but in some cases it does not and it has the worst asymptotic performance.

5. Discussion

It was observed in the simulations of Marron et al. [10], and it was proved in terms of probabilities of misclassification of new data points in Hall et al. [7], that in some cases the MD, SVM and DWD methods have asymptotically the same behavior as d tends to infinity, for fixed sample sizes and under a Gaussian assumption. Considering two classes of Gaussian data with common diagonal covariance matrix, one class with mean zero and the other with mean v_d, in this paper we provide new theoretical explanations of these results in terms of the angles between the normal vectors of the separating hyperplanes and the optimal direction v_d. In particular, we show that the MD, SVM, DWD and MDP methods have asymptotically the same behavior as d tends to infinity. This result is given by Theorem 3.1. It is observed in this result that the asymptotic consistency and strong inconsistency of these methods depend on the asymptotic behavior of $\|v_d\|$ as the dimension increases.

Furthermore, in this paper we present a comparison between the NB method with the other four methods, not treated before in [7] nor [10], only a comparison by simulations is done in [1]. We observe that this method sometimes behaves worse than the other methods. This result follows from Proposition 3.1 where we show that the consistency of the NB method is driven by properties of the means of the classes.

As pointed out by the referees, another method that can be considered to compare with those presented here is Regularized Logistic Regression (RLR) of le Cessie and van Houwelingen [9]. By simulations in terms of proportion of misclassification of new data points, a comparison of SVM, DWD and MD with RLR is done in [10]. It is seen in their Fig. 2(a), which corresponds to Gaussian data with identity covariance matrix, that the MD and DWD methods have the best performance, however all the methods tend to have the same behavior as dimension increases. Then, it is an interesting open problem to know if Theorem 3.1 holds for RLR. The difficulty of this problem is that the normal vector of the separating hyperplane of this method does not have a closed form, this makes complicated the study of the angle between the optimal direction v_d and the normal vector of this method. On the other hand, by the results of [10] it is expected that the MD method has better behavior than RLR.

6. Proofs of results

Proof of Lemma 3.1. Let $z_i = \Sigma^{-1/2}(x_i - v_d)$ for $i = 1, 2, \ldots, m$; $y_j = x_{m+j}$ and $w_j = \Sigma^{-1/2}y_j$ for $j = 1, 2, \ldots, n$. Then the z_i's and w_j's are independent random vectors with d-multivariate standard normal distribution. By Kolmogorov's strong law for independent random variables (see [14, Theorem 2.3.10]), the assumption (30) and since the sequence $\{\sigma_k\}_{k=1}^{\infty}$ is bounded we have the following:

$$\begin{aligned}
&\frac{\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)2}}{d} \xrightarrow{w} \sigma^2, \qquad \frac{\sum_{k=1}^{d}\sigma_k^2 w_j^{(k)2}}{d} \xrightarrow{w} \sigma^2, \qquad \frac{\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)} w_j^{(k)}}{d} \xrightarrow{w} 0, \quad \forall i, \forall j,\\
&\frac{\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)} z_j^{(k)}}{d} \xrightarrow{w} 0, \qquad \frac{\sum_{k=1}^{d}\sigma_k^2 w_i^{(k)} w_j^{(k)}}{d} \xrightarrow{w} 0, \quad i \neq j,
\end{aligned} \tag{36}$$

as $d \to \infty$. Observe that for $i \neq j$

$$\begin{aligned}
\frac{\|x_i - x_j\|^2}{d} &= \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2\left(\frac{x_i^{(k)} - v_d^{(k)}}{\sigma_k} - \frac{x_j^{(k)} - v_d^{(k)}}{\sigma_k}\right)^2 = \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2(z_i^{(k)} - z_j^{(k)})^2\\
&= \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)2} - \frac{2}{d}\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)} z_j^{(k)} + \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2 z_j^{(k)2},
\end{aligned}$$

then by (36) we have

$$\frac{\|x_i - x_j\|}{d^{1/2}} \xrightarrow{P} 2^{1/2}\sigma \quad \text{as } d \to \infty. \tag{37}$$

Analogously, for $i \neq j$

$$\frac{\|y_i - y_j\|}{d^{1/2}} \xrightarrow{P} 2^{1/2}\sigma \quad \text{as } d \to \infty. \tag{38}$$

We also have that $\forall i, j$

$$\frac{\|x_i - y_j\|^2}{d} = \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2\left(\frac{x_i^{(k)} - v_d^{(k)}}{\sigma_k} - \frac{y_j^{(k)}}{\sigma_k} + \frac{v_d^{(k)}}{\sigma_k}\right)^2 = \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2\left(z_i^k - w_j^{(k)} + \frac{v_d^{(k)}}{\sigma_k}\right)^2$$
$$= \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2(z_i^{(k)} - w_j^{(k)})^2 + \frac{2}{d}\sum_{k=1}^{d}\sigma_k v_d^{(k)}(z_i^{(k)} - w_j^{(k)}) + \frac{\|v_d\|^2}{d}. \tag{39}$$

From (36) the first term of (39) converges in distribution to $2\sigma^2$ as $d \to \infty$. Observe that the second term of (39) is equal in distribution to

$$2^{3/2}\left(\frac{\sum_{k=1}^{d}\sigma_k^2 v_d^{(k)2}}{d}\right)^{1/2}\frac{N_0}{d^{1/2}}, \tag{40}$$

where N_0 is a random variable with standard normal distribution. By hypothesis $\{\sigma_k\}_{k=1}^{\infty}$ is bounded by some constant $M > 0$, then

$$\frac{\sum_{k=1}^{d}\sigma_k^2 v_d^{(k)2}}{d} \le M^2\frac{\|v_d\|^2}{d}. \tag{41}$$

Furthermore, the sequence $\{\|v_d\|d^{-1/2}\}_{d=1}^{\infty}$ is bounded by some $R > 0$, since $\|v_d\|d^{-1/2} \to c$ as $d \to \infty$ with $0 \le c < \infty$. Thus the left side of (41) is bounded by M^2R^2 and (40) converges in distribution to zero as $d \to \infty$. Hence, from (39) we have

$$\frac{\|x_i - y_j\|}{d^{1/2}} \xrightarrow{P} \ell = (2\sigma^2 + c^2)^{1/2} \quad \text{as } d \to \infty. \tag{42}$$

Note that (37), (38) and (42) imply that the data $x_1, \ldots, x_N$ tend to be the vertices of an N-polyhedron (a figure in $(N-1)$-dimensional space with just N vertices and all faces given by $(N-2)$-dimensional hyperplanes) as $d \to \infty$. This polyhedron has m of its vertices arranged as those of an m-simplex (an m-polyhedron with all edges of equal length) 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 $2^{1/2}$. 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.

If $\widetilde{v} = X_+\alpha_+ - X_-\alpha_-$ is proportional to the SVM normal vector, as seen in Section 2.1, $\widetilde{v}$ is the difference between the two closest vectors of the convex hulls of the classes C_+ and C_-. When $d \to \infty$ these convex hulls tend to be the m-simplex and n-simplex, respectively. We will show that the closest points of these simplices are the means $\bar{x}^* = \sum_{i=1}^{m} x_i^*/m$ and $\bar{y}^* = \sum_{i=1}^{n} y_i^*/n$. For the N-polyhedron we have

$$\|x_i^* - y_j^*\| = \ell,$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Since the distance from x_i^* to any vertex of the n-simplex is the same, we have that for any j, x_i^*, y_j^* and $\bar{y}^*$ are the vertices of a right-angled triangle where the hypotenuse is the line joining x_i^* to y_j^*. Thus $\bar{y}^*$ is the closest point in the n-simplex to x_i^*, for $i = 1, 2, \ldots, m$. Similarly, because the distance from $\bar{y}^*$ to x_i^* for $i = 1, 2, \ldots, m$ is constant, the closest point in the m-simplex to $\bar{y}^*$ is $\bar{x}^*$, hence the closest points in the simplices are $\bar{x}^*$ and $\bar{y}^*$. Thus we have (31) in the SVM case.

Now for DWD, in the case when $\widetilde{v}$ is proportional to the DWD normal vector, by Section 2.2 α solves the optimization problem (17). For the N-polyhedron we will show that this α is given by

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

This is a consequence of the fact that if β satisfies $\mathbf{1}_m^\top\beta_+ = \mathbf{1}_n^\top\beta_+ = 1$ then

$$\|X_+^*\beta_+ - X_-^*\beta_-\| \ge \|\bar{x}^* - \bar{y}^*\| = \|X_+^*\widehat{\alpha}_+ - X_-^*\widehat{\alpha}_-\|$$

where $X_+^* = [x_1^*, \ldots, x_m^*]$ and $X_-^* = [y_1^*, \ldots, y_n^*]$, since $\bar{x}^*$ and $\bar{y}^*$ are the closest points of the simplices. Furthermore

$$(\mathbf{1}_m^\top\sqrt{\beta_+} - \mathbf{1}_n^\top\sqrt{\beta_-})^2 \le (\sqrt{m} + \sqrt{n})^2 = (\mathbf{1}_m\sqrt{\widehat{\alpha}_+} + \mathbf{1}_n\sqrt{\widehat{\alpha}_-})^2,$$

thus

$$\frac{(\mathbf{1}_m\sqrt{\beta_+}+\mathbf{1}_n\sqrt{\beta_-})^2}{\|X_+^*\beta_+ - X_-^*\beta_-\|} \le \frac{(\mathbf{1}_m\sqrt{\widehat{\alpha}_+}+\mathbf{1}_n\sqrt{\widehat{\alpha}_-})^2}{\|X_+^*\widehat{\alpha}_+ - X_-^*\widehat{\alpha}_-\|},$$

and $\widehat{\alpha}$ solves the optimization problem (17) for the N-polyhedron, hence we have (31).

For the case of the MD method $\widetilde{v} = \bar{x}_+ - \bar{x}_-$, thus

$$\alpha_{i,+} = \frac{1}{m}, \qquad \alpha_{j,-} = \frac{1}{n}, \quad i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n.$$

Therefore we have (31). □

Proof of Theorem 3.1. *Case* 1: *When v is the normal vector of the MD, SVM or DWD hyperplane*. We have seen in Section 2 that v is proportional to the vector $\widetilde{v} = X_+\alpha_+ - X_-\alpha_-$ given in Lemma 3.1. We also have

$$\cos(\text{Angle}(\widetilde{v}, v_d)) = \frac{\langle \widetilde{v}, v_d\rangle}{\|\widetilde{v}\|\,\|v_d\|}.$$

Let z_i, y_j and w_j, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, be as in the proof of Lemma 3.1. It can be seen that

$$\begin{aligned}\langle \widetilde{v}, v_d\rangle &= \sum_{k=1}^{d}\sigma_k v_d^{(k)}\left(\sum_{i=1}^{m}\alpha_{i,+}z_i^{(k)} - \sum_{i=1}^{n}\alpha_{i,-}w_i^{(k)}\right) + \|v_d\|^2\\ &= \sum_{i=1}^{m}\alpha_{i,+}\sum_{k=1}^{d}\sigma_k v_d^{(k)}z_i^{(k)} - \sum_{i=1}^{n}\alpha_{i,-}\sum_{k=1}^{d}\sigma_k v_d^{(k)}w_i^{(k)} + \|v_d\|^2.\end{aligned}$$

Therefore

$$\langle \widetilde{v}, v_d\rangle = \left(\sum_{k=1}^{d}\sigma_k^2 v_d^{(k)2}\right)^{1/2}\left(\sum_{i=1}^{m}\alpha_{i,+}N_{i,+} - \sum_{i=1}^{n}\alpha_{i,-}N_{i,-}\right) + \|v_d\|^2, \tag{43}$$

where

$$N_{i,+} = \left(\sum_{k=1}^{d}\sigma_k^2 v_d^{(k)2}\right)^{-1/2}\sum_{k=1}^{d}\sigma_k v_d^{(k)}z_i^{(k)} \quad \text{and} \quad N_{j,-} = \left(\sum_{k=1}^{d}\sigma_k^2 v_d^{(k)2}\right)^{-1/2}\sum_{k=1}^{d}\sigma_k v_d^{(k)}w_j^{(k)},$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Observe that the $N_{i,+}$'s and $N_{i,-}$'s are independent random variables with the standard normal distribution. Furthermore

$$\begin{aligned}\|\widetilde{v}\|^2 &= \sum_{k=1}^{d}\left(\sum_{i=1}^{m}\sigma_k\alpha_{i,+}z_i^{(k)} - \sum_{i=1}^{n}\sigma_k\alpha_{i,-}w_i^{(k)} + v_d^{(k)2}\right)^2\\ &= \sum_{i=1}^{m}\alpha_{i,+}^2\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)2} + 2\sum_{i<j}\alpha_{i,+}\alpha_{j,+}\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)}z_j^{(k)} + \sum_{i=1}^{n}\alpha_{i,-}^2\sum_{k=1}^{d}\sigma_k^2 w_i^{(k)2}\\ &\quad + 2\sum_{i<j}\alpha_{i,-}\alpha_{j,-}\sum_{k=1}^{d}\sigma_k^2 w_i^{(k)}w_j^{(k)} - 2\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{i,+}\alpha_{j,-}\sum_{k=1}^{d}\sigma_k^2 z_i^{(k)}w_j^{(k)}\\ &\quad + 2\left(\sum_{k=1}^{d}\sigma_k^2 v_d^{(k)2}\right)^{1/2}\left(\sum_{i=1}^{m}\alpha_{i,+}N_{i,+} - \sum_{i=1}^{n}\alpha_{i,-}N_{i,-}\right) + \|v_d\|^2.\end{aligned} \tag{44}$$

Note that

$$\frac{N_{i,+}}{d^{1/2}} \xrightarrow{w} 0, \qquad \frac{N_{j,-}}{d^{1/2}} \xrightarrow{w} 0 \quad \text{as } d \to \infty. \tag{45}$$

For the case when $\|v_d\|d^{-1/2} \to \infty$, from (36), (41), (43)–(45) and since $0 \le \alpha_{i,+}, \alpha_{j,-} \le 1$ we have

$$\frac{\|\widetilde{v}\|^2}{\|v_d\|^2} \xrightarrow{w} 1, \qquad \frac{\langle \widetilde{v}, v_d\rangle}{\|v_d\|^2} \xrightarrow{w} 1, \tag{46}$$

as $d \to \infty$. Thus

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

and

$$\text{Angle}(\widetilde{v}, v_d) = \arccos\left(\frac{\langle \widetilde{v}, v_d\rangle}{\|\widetilde{v}\|\,\|v_d\|}\right) \xrightarrow{w} 0,$$

as $d \to \infty$.

For the case when $\|v_d\|d^{-1/2} \to c$, with $0 \le c < \infty$, from (36), (41), (43)–(45) and Lemma 3.1 it can be seen that

$$\frac{\|\widetilde{v}\|^2}{d} \xrightarrow{w} \gamma\sigma^2 + c^2, \qquad \frac{\langle \widetilde{v}, v_d\rangle}{d^{1/2}\|v_d\|} \xrightarrow{w} c, \tag{47}$$

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

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

and

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

as $d \to \infty$. In particular, for $c = 0$ we have $\arccos(c/(\gamma\sigma^2 + c^2)^{1/2}) = \pi/2$.

Case 2: When v is the normal vector of the MDP hyperplane. Let $\bar{x} = \sum_{i=1}^m x_i/m$, $\bar{z} = \sum_{i=1}^m z_i/m$, $\bar{y} = \sum_{i=1}^n y_i/n$, $\overline{w} = \sum_{i=1}^n w_i/n$, where the z_i's, y_j's and w_j's are as in the proof of Case 1. Note that

$$\begin{aligned}
\|x_i - \bar{x}\|^2 &= \sum_{k=1}^d \sigma_k^2 (z_i^{(k)} - \bar{z}^{(k)})^2 = \sum_{k=1}^d \sigma_k^2 \left[\left(1 - \frac{1}{m}\right) z_i^{(k)} - \frac{1}{m}\sum_{j\neq i} z_j^{(k)}\right]^2 \\
&= \left(1 - \frac{1}{m}\right)^2 \sum_{k=1}^d \sigma_k^2 z_i^{(k)2} - 2\left(1 - \frac{1}{m}\right)\sum_{k=1}^d \sigma_k^2 z_i^{(k)} \sum_{j\neq i} \frac{z_j^{(k)}}{m} + \sum_{k=1}^d \sigma_k^2 \left(\sum_{j\neq i} \frac{z_j^{(k)}}{m}\right)^2 \\
&= \left(1 - \frac{1}{m}\right)^2 \sum_{k=1}^d \sigma_k^2 z_i^{(k)2} - 2\left(1 - \frac{1}{m}\right)\frac{(m-1)^{1/2}}{m} \sum_{k=1}^d \sigma_k^2 z_i^{(k)} N_{-i}^{(k)} + \frac{m-1}{m^2}\sum_{k=1}^d \sigma_k^2 N_{-i}^{(k)2}
\end{aligned} \tag{48}$$

where the $N_{-i}^{(k)} = (m-1)^{-1/2}\sum_{j\neq i} z_j^{(k)}$, for $k = 1, 2, \ldots, d$, are independent random variables with the standard normal distribution that are independent of the z_i. Let $u = \bar{x} - \bar{y}$, then

$$\begin{aligned}
\langle x_i - \bar{x}, u\rangle &= \sum_{k=1}^d \sigma_k^2 (z_i^{(k)} - \bar{z}^{(k)}) \left(\bar{z}^{(k)} - \overline{w}^{(k)} + \frac{v_d^{(k)}}{\sigma_k}\right) \\
&= \frac{1}{m}\sum_{k=1}^d \sigma_k^2 z_i^{(k)2} + \sum_{k=1}^d \sigma_k^2 z_i^{(k)} \left(\sum_{j\neq i} \frac{z_j^{(k)}}{m} - \sum_{j=1}^n \frac{w_j^{(k)}}{n}\right) - \sum_{k=1}^d \sigma_k^2 \bar{z}^{(k)2} \\
&\quad + \sum_{k=1}^d \sigma_k^2 \bar{z}^{(k)} \overline{w}^{(k)} + \sum_{k=1}^d \sigma_k v_d^{(k)} z_i^{(k)} - \sum_{k=1}^d \sigma_k v_d^{(k)} \bar{z}^{(k)}.
\end{aligned} \tag{49}$$

If $\|v_d\|d^{-1/2} \to \infty$ using (41), (48), (49) and Kolmogorov's strong law we can see that

$$\frac{\|x_i - \bar{x}\|^2}{d} \xrightarrow{w} \frac{m-1}{m}\sigma^2, \qquad \frac{\langle x_i - \bar{x}, u\rangle}{d^{1/2}\|v_d\|} \xrightarrow{w} 0$$

as $d \to \infty$. By (46) we have that $\|u\|/\|v_d\| \xrightarrow{w} 1$ as $d \to \infty$. Then

$$\cos(\text{Angle}(x_i - \bar{x}, u)) = \frac{\langle x_i - \bar{x}, u\rangle}{\|x_i - \bar{x}\|\,\|u\|} = \left(\frac{d^{1/2}}{\|x_i - \bar{x}\|}\right)\left(\frac{\|v_d\|}{\|u\|}\right)\left(\frac{\langle x_i - \bar{x}, u\rangle}{d^{1/2}\|v_d\|}\right) \xrightarrow{w} 0$$

as $d \to \infty$. Analogously, $\cos(\text{Angle}(y_i - \bar{y}, u)) \xrightarrow{w} 0$ as $d \to \infty$. Thus

$$\text{Angle}(x_i - \bar{x}, u) \xrightarrow{w} \frac{\pi}{2}, \qquad \text{Angle}(y_j - \bar{y}, u) \xrightarrow{w} \frac{\pi}{2} \tag{50}$$

as $d \to \infty$. The same is true if $\|v_d\|d^{-1/2} \to c$, with $c \ge 0$.

Let C be the matrix whose columns are the vectors $x_i-\bar{x}$, $y_j-\bar{y}$, for $i=1,2,\ldots,m$ and $j=1,2,\ldots,n$. By Section 2.4 the normal vector of the MDP method is given by $v=Qu/\|Qu\|$ where Q is the symmetric projection matrix on the orthogonal complement of the column space of C. According to (50), u tends to be in the orthogonal complement of the column space of C. Thus when d is large Qu can be approximated by u, and v can be approximated by $u/\|u\|$. Therefore

$$\cos(\text{Angle}(v,v_d))=\frac{\langle v,v_d\rangle}{\|v\|\,\|v_d\|} \tag{51}$$

can be approximated by $\langle u,v_d\rangle/(\|u\|\,\|v_d\|)$. Hence by Case 1, it converges to 1 if $\|v_d\|d^{-1/2}\to\infty$ and converges to $c/(\gamma\sigma^2+c^2)^{1/2}$ if $\|v_d\|d^{-1/2}\to c$ with $0\le c<\infty$. □

Proof of Proposition 3.1. *Proof of (a):* Let $z_i, y_j, w_j, \bar{x}, \bar{z}, \bar{y}$ and $\bar{w}$ be as in the proof of Theorem 3.1. We observe that $\|v_d\|=\beta d^{1/2}$ and therefore $\|v_d\|d^{-1/2}=\beta\to c$ as $d\to\infty$. Let $\widehat{s}_1,\ldots,\widehat{s}_d$ be the diagonal entries of the pooled covariance matrix $\widehat{\Sigma}$ given by (26). They are independent random variables with distribution $\sigma_k^2\chi^2_{(m+n-2)}/(m+n-2)$ since the x_i's are independent of the y_j's, for $i=1,2,\ldots,m, j=1,2,\ldots,n$. The NB normal vector given by (25) is equal to

$$v=D^{-1}(\bar{x}-\bar{y})=((\bar{x}^{(1)}-\bar{y}^{(1)})/\widehat{s}_1,\ldots,(\bar{x}^{(d)}-\bar{y}^{(d)})/\widehat{s}_d)^\top, \tag{52}$$

where $D=\text{diag}(\widehat{\Sigma})=\text{diag}(\widehat{s}_1,\ldots,\widehat{s}_d)$. Let $\widetilde{s}_k=\widehat{s}_k/\sigma_k^2$, then we have that

$$\begin{aligned}\frac{\|v\|^2}{d}&=\sum_{k=1}^{d}\frac{(\bar{x}^{(k)}-\bar{y}^{(k)})^2}{d\widehat{s}_k^2}=\sum_{k=1}^{d}\sigma_k^2\frac{(\bar{z}^{(k)}-\bar{w}^{(k)}+\beta/\sigma_k)^2}{d\widehat{s}_k^2}\\&=\sum_{k=1}^{d}\frac{(\bar{z}^{(k)}-\bar{w}^{(k)})^2}{d\sigma_k^2\widetilde{s}_k^2}+2\beta\sum_{k=1}^{d}\frac{\bar{z}^{(k)}-\bar{w}^{k}}{d\sigma_k^3\widetilde{s}_k^2}+\beta^2\sum_{k=1}^{d}\frac{1}{d\sigma_k^4\widetilde{s}_k^2}\end{aligned} \tag{53}$$

and

$$\begin{aligned}\cos(\text{Angle}(v,v_d))&=\frac{\langle v,v_d\rangle}{\|v\|\,\|v_d\|}=\frac{1}{\|v\|\,\|v_d\|}\sum_{k=1}^{d}\frac{(\bar{x}^{(k)}-\bar{y}^{(k)})}{\widehat{s}_k}v_d^{(k)}\\&=\frac{d^{1/2}}{\|v\|}\left[\sum_{k=1}^{d}\frac{\bar{z}^{(k)}-\bar{w}^{(k)}}{d\sigma_k\widetilde{s}_k}+\beta\sum_{k=1}^{d}\frac{1}{d\sigma_k^2\widetilde{s}_k}\right].\end{aligned} \tag{54}$$

Observe that the random variables $(\bar{x}^{(k)}-\bar{y}^{(k)})$'s are independent of the $\widehat{s}_k$'s, since the x_i's are independent of the y_j's and in the normal case the sample mean is independent of the sample variance. Thus the $(\bar{z}^{(k)}-\bar{w}^{(k)})$'s are independent of the $\widehat{s}_k$'s and $\widetilde{s}_k$'s. Let $\widetilde{\mu}=E(1/\widetilde{s}_k)$ and $\widetilde{\sigma}^2=\text{Var}(1/\widetilde{s}_k)$. Since $1/\widetilde{s}_k$ is a multiple of a random variable with the *inverse-chi-square distribution* with $m+n-2$ degrees of freedom, $\widetilde{\mu}$ and $\widetilde{\sigma}^2$ are given by (34) and are finite because $m+n>6$. Let $\gamma=1/m+1/n$.

Since $R<\sigma_k<M\ \forall k=1,2,\ldots$ for some $R, M>0$ and we have (32), by Kolmogorov's strong law, (53) and (54) it follows that if $0\le c<\infty$ then

$$\frac{\|v\|^2}{d}\xrightarrow{w}(\gamma\rho_1+c^2\rho_2)(\widetilde{\sigma}^2+\widetilde{\mu}^2) \tag{55}$$

and

$$\cos(\text{Angle}(v,v_d))\xrightarrow{w}\frac{c\widetilde{\mu}\rho_1}{(\gamma\rho_1+c^2\rho_2)^{1/2}(\widetilde{\sigma}^2+\widetilde{\mu}^2)^{1/2}} \tag{56}$$

as $d\to\infty$. Analogously, if $c=\infty$ from (53) and (54) it follows that

$$\frac{\|v\|^2}{d\beta^2}\xrightarrow{w}(\widetilde{\sigma}^2+\widetilde{\mu}^2)\rho_2 \tag{57}$$

and

$$\cos(\text{Angle}(v,v_d))\xrightarrow{w}\frac{\widetilde{\mu}}{(\widetilde{\sigma}^2+\widetilde{\mu}^2)^{1/2}\rho_2^{1/2}} \tag{58}$$

as $d\to\infty$. Applying the arccosine function in (56) and (58) we get (33).

Proof of (b): In this case $\|v_d\| = d^\delta$. We have

$$\|v\|^2 = \sum_{k=1}^{d} \frac{(\bar{x}^{(k)} - \bar{y}^{(k)})^2}{\widehat{s}_k^2} = \frac{\sigma_1^2(\bar{z}^{(1)} - \bar{w}^{(1)} + d^\delta/\sigma_1)^2}{\widehat{s}_1^2} + \sum_{k=2}^{d} \frac{\sigma_k^2(\bar{z}^{(k)} - \bar{w}^{(k)})^2}{\widehat{s}_k^2}$$
$$= \frac{\sigma_1^2(\bar{z}^{(1)} - \bar{w}^{(1)})^2}{\widehat{s}_1^2} + \frac{2\sigma_1^2(\bar{z}^{(1)} - \bar{w}^{(1)})d^\delta}{\widehat{s}_1^2} + \frac{d^{2\delta}}{\widehat{s}_1^2} + \sum_{k=2}^{d} \frac{\sigma_k^2(\bar{z}^{(k)} - \bar{w}^{(k)})^2}{\widehat{s}_k^2}, \tag{59}$$

and

$$\cos(\text{Angle}(v, v_d)) = \frac{\langle v, v_d\rangle}{\|v\|\,\|v_d\|} = \frac{(\bar{x}^{(1)} - \bar{y}^{(1)})}{\|v\|\widehat{s}_1}$$
$$= \frac{\sigma_1(\bar{z}^{(1)} - \bar{w}^{(1)})}{\|v\|\widehat{s}_1} + \frac{d^\delta}{\|v\|\widehat{s}_1}. \tag{60}$$

If $\delta < 1/2$, by Kolmogorov's strong law, (59) and (60) it follows that

$$\frac{\|v\|^2}{d} \xrightarrow{w} \gamma(\widetilde{\sigma}^2 + \widetilde{\mu}^2)\rho_1$$

and

$$\cos(\text{Angle}(v, v_d)) = \frac{d^{1/2}}{\|v\|}\frac{(\bar{z}^{(1)} - \bar{w}^{(1)})}{d^{1/2}\sigma_1\widetilde{s}_1} + \frac{d^{1/2}}{\|v\|}\frac{1}{d^{1/2-\delta}\sigma_1^2\widetilde{s}_1} \xrightarrow{w} 0 \tag{61}$$

as $d \to \infty$. Analogously, if $\delta > 1/2$ we have that

$$\frac{\|v\|^2\widehat{s}_1^2}{d^{2\delta}} \xrightarrow{w} 1$$

and

$$\cos(\text{Angle}(v, v_d)) = \frac{d^\delta}{\|v\|\widehat{s}_1}\frac{\sigma_1(\bar{z}^{(1)} - \bar{w}^{(1)})}{d^\delta} + \frac{d^\delta}{\|v\|\widehat{s}_1} \xrightarrow{w} 1 \tag{62}$$

as $d \to \infty$. For the case $\delta = 1/2$ we have

$$\frac{\|v\|^2\widehat{s}_1^2}{d} \xrightarrow{w} 1 + \gamma(\widetilde{\sigma}^2 + \widetilde{\mu}^2)\rho_1\sigma_1^4\widetilde{s}_1^2$$

and

$$\cos(\text{Angle}(v, v_d)) = \frac{d^{1/2}}{\|v\|\widehat{s}_1}\frac{\sigma_1(\bar{z}^{(1)} - \bar{w}^{(1)})}{d^{1/2}} + \frac{d^{1/2}}{\|v\|\widehat{s}_1} \xrightarrow{w} [1 + \gamma(\widetilde{\sigma}^2 + \widetilde{\mu}^2)\rho_1\sigma_1^4\widetilde{s}_1^2]^{-1/2}, \tag{63}$$

as $d \to \infty$. Applying the arccosine function to (61)–(63) we get (35). □

Acknowledgments

The authors are grateful to the referees for their valuable suggestions and comments to improve this paper. They also thank Victor Perez-Abreu for his suggestions on an early version of it. They thank the Centro de Investigacion en Matematicas (CIMAT) and the Statistical and Applied Mathematical Sciences Institute (SAMSI) for the support provided during the elaboration of this paper. The first author also thanks CONACYT for her scholarship for Ph.D. study in CIMAT (2007–2010), and for her scholarship to do an academic visit to the UNC at Chapel Hill (August–December 2009).

Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2012.10.001.

References

[1] J. Ahn, J.S. Marron, The maximal data piling direction for discrimination, Biometrika 97 (1) (2010) 254–259.
[2] P.J. Bickel, E. Levina, Some theory for Fisher's linear discriminant function, "Naive Bayes" and some alternatives where there are many more variables than observations, Bernoulli 10 (2004) 989–1010.

[3] A. Bolivar-Cime, J.S. Marron, Supplementary material for "comparison of binary discrimination methods for high dimension low sample size data", September 2012. Manuscript Available at: https://sites.google.com/site/addybolivarcime/publications/supplementary-files.
[4] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.
[5] C. Cortes, V.N. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[6] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.
[7] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Stat. Soc. Ser. B 67 (3) (2005) 427–444.
[8] A.J. Izenman, Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning, in: Springer Texts in Statistics, Springer, New York, 2008.
[9] S. le Cessie, J.C. van Houwelingen, Ridge estimators in logistic regression, Appl. Statist. 41 (1) (1992) 191–201.
[10] J.S. Marron, M.J. Todd, J. Ahn, Distance-weighted discrimination, J. Amer. Statist. Assoc. 102 (480) (2007) 1267–1271.
[11] J.S. Marron, M.J. Todd, J. Ahn, Distance-weighted discrimination, February 2007. Manuscript Available at: http://www.stat.uga.edu/~jyahn/DWD/.
[12] X. Qiao, H.H. Zhang, Y. Liu, M.J. Todd, J.S. Marron, Weighted distance weighted discrimination and its asymptotic properties, J. Amer. Statist. Assoc. 105 (489) (2010) 401–414.
[13] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, The MIT Press, Cambridge, Massachusetts, 2002.
[14] P.K. Sen, J.M. Singer, Large Sample Methods in Statistics, Chapman & Hall, Inc., New York, 1993.
[15] V.N. Vapnik, Estimation of Dependences Based on Empirical Data, in: Springer Series in Statistics, Springer-Verlag, Berlin, 1982.
[16] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, 1995.