Asymptotic Properties of Distance-Weighted Discrimination

Xingye Qiao[1], Hao Helen Zhang[2], Yufeng Liu[13], Michael J. Todd[4], J. S. Marron[1]

September 3, 2008

Abstract

While Distance-Weighted Discrimination (DWD) is an appealing approach to classification in high dimensions, it was designed for balanced data sets. In the case of unequal costs, biased sampling or unbalanced data, there are major improvements available, using appropriately weighted versions of DWD. A major contribution of this paper is the development of optimal weighting schemes for various nonstandard classification problems. The second major contribution is substantial asymptotic study of both the original and the weighted DWD. Let n be the sample size and d be the dimension of data. Both conventional (n-asymptotic) Fisher consistency and high dimension low sample size asymptotics (d-asymptotics) are studied.

Key Words and Phrases: High dimension, low sample size data; Linear discrimination; Fisher Consistency; Non-standard asymptotics; Unbalanced data.

[1]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695

[3]Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, North Carolina 27599

[4]School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853

1 Introduction

The Support Vector Machine (SVM) (Vapnik 1995; Cristianini et al. 2000; Duda, Hart and Stork 2001; Schölkopf and Smola 2002) is a powerful tool for classification in machine learning statistical research. It is a well known example of a large margin classifier. Distance-Weighted Discrimination (DWD) is a more recent large margin classifier specifically designed for high-dimension, low-sample-size (HDLSS) data settings, where the dimension d is far greater than the sample size n. As noted in Marron et al. (2007), DWD can resolve the *data-piling* problem that is often observed for the SVM in HDLSS settings.

The main idea behind SVM in the separable case is to find the separating hyperplane maximizing the distance of the hyperplane from the closest data point of each class. Although the SVM gives good performance in many real applications, it may suffer from a loss of generalizability due to data-piling in HDLSS settings, i.e., the support vectors tend to pile at the boundaries of the margin when projected towards the separating hyperplanes. This can adversely affect the generalization performance of the SVM (Marron et al. 2007). To overcome this data-piling issue, Marron et al. (2007) proposed DWD, which finds the hyperplane minimizing the sum of the reciprocals r_i of the distances from every data point to the margin ($\min \sum_i r_i^{-1}$).

Although the standard DWD effectively avoids the data-piling problem in HDLSS settings, its decision boundary heavily depends on the ratio of sample sizes from two classes. In other words, it treats the reciprocal distances from the two classes as equally important. As a consequence, a larger sample size for the majority class will push the separating hyperplane toward the minority class because it will make $\sum_i r_i^{-1}$ small. This can result in inefficient generalization properties. A straightforward method to improve the performance of the standard DWD is to allow different weights on these two classes, with greater weight on the minority class and smaller weight on the majority class. We call this *weighted DWD*. Qiao and Liu (2008) proposed an adaptive choice of weights for general classification methods. Section

2 provides the globally optimal weighting schemes under a variety of situations and proves Fisher consistency of weighted DWD, using classical asymptotic methods (where $n \to \infty$).

A second issue of great interest is the asymptotic properties of the DWD classifier in HDLSS settings. Ge and Simpson (1998) analyzed the high-dimensional asymptotics of some classical discriminant analysis, emphasizing the effects of correlations. For our purpose, the special structure of multi-class data in high-dimensional situations is very useful. Hall et al. (2005) showed that there exists a *geometric representation* of data in the high-dimensional case, under the condition that the d variables of each observation are almost independent. The main result of Hall et al. (2005) is that, under appropriate conditions, pairwise distances between the n^+ (or n^-) data points within class $+1$ (or -1) are approximately constant as $d \to \infty$ with n^+ (or n^-) fixed. As a consequence, each data set of one class can be viewed as an n^+ (or n^-)-simplex within the d-dimensional space. Ahn et al. (2007) extended this work by showing that the conditions can be milder. However, a counterexample discussed in detail in Jung and Marron (2008) suggests that an additional assumption is needed. In this article, our theory is based on an additional condition, which is still milder than that in Hall et al. (2005).

The next step is to give a geometric representation of the data set from two classes under these milder conditions. Once the data framework is built in Section 3, we study the asymptotic properties of two aspects of DWD as $d \to \infty$. Both properties are tightly related to the geometric representation described above for high-dimensional situations. The first aspect is the classification error rate of DWD. The second is the similarity of the DWD direction to the optimal linear classification direction. Both aspects are driven by appropriate notions of signal to noise ratios, defined in terms of class means and intraclass variances.

In this paper, we propose optimal weighting schemes in Section 2 for each situation (unequal cost, biased sampling). With the presence of unbalanced data, we specially find the corresponding weighting scheme under the *Mean Within Group Error criterion* proposed

by Qiao and Liu (2008). Each weighting scheme can be shown to be Fisher consistent as $n \to \infty$. In order to discuss the HDLSS asymptotic properties, in Section 3, we represent the data samples from two classes geometrically in the high-dimensional case. With the help of this representation, we give the d-asymptotic properties of DWD. A numerical study is done in Section 4 based on simulated data in both low-dimensional and high-dimensional cases, and real data in high-dimensional situations. Conclusions are given in Section 5. Proofs of the Propositions and Theorems are included in the Appendix.

2 Weighted DWD

2.1 General Classification Problems

Consider the problem of classifying a subject associated with vector $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ into one of two classes with the label $Y \in \{\pm 1\}$. Assume the target population has an unknown probability distribution $P(\mathbf{X}, Y)$, and the examples are independently and identically generated from $P(\mathbf{X}, Y)$. Let the unconditional (prior) probabilities of the populations be $\pi^+ = \mathrm{P}(Y = +1)$ and $\pi^- = \mathrm{P}(Y = -1)$ and let $g^+(\mathbf{x})$ and $g^-(\mathbf{x})$ denote the conditional densities given $Y = +1$ and $Y = -1$ respectively. Then the conditional (posterior) probability of a subject belonging to the "+1" class given $\mathbf{X} = \mathbf{x}$ is

$$p(\mathbf{x}) = \mathrm{Prob}(Y = +1|\mathbf{X} = \mathbf{x}) = \frac{\pi^+ g^+(\mathbf{x})}{\pi^+ g^+(\mathbf{x}) + \pi^- g^-(\mathbf{x})}. \tag{1}$$

2.1.1 Unequal costs

Assume different costs are used for different types of misclassification, say, classifying a "+1" subject as "−1" represents a more serious error than classifying a "−1" subject as "+1". As an example, failing to diagnose a potentially fatal illness may be viewed as substantially more *costly* than concluding that the disease is present when, in fact, it is not. We will use c^+ for the false-positive cost and c^- for the false-negative cost. Table 1 shows these costs.

Using the overall misclassification criterion, for any classifier ϕ, where either $\phi(\mathbf{x}) = +1$ or $\phi(\mathbf{x}) = -1$, its loss function for classifying a pair $(\mathbf{x}, y)$ is defined as $L[\phi] = c^+ I[y = -1]I[\phi(\mathbf{x}) = +1] + c^- I[y = +1]I[\phi(\mathbf{x}) = -1]$. Given $\mathbf{x}$, the risk, i.e. the expected loss of ϕ is

$$\begin{aligned} E[L(\phi)|\mathbf{X} = \mathbf{x}] &= c^+ \mathrm{P}(Y = -1|\mathbf{X} = \mathbf{x})I[\phi(\mathbf{x}) = +1] + c^- \mathrm{P}(Y = +1|\mathbf{X} = \mathbf{x})I[\phi(\mathbf{x}) = -1] \\ &= c^+[1 - p(\mathbf{x})]I[\phi(\mathbf{x}) = +1] + c^- p(\mathbf{x})I[\phi(\mathbf{x}) = -1]. \end{aligned}$$

		Classify as	
		$+1$	-1
True population:	$+1$	0	c^-
	-1	c^+	0

Table 1: Unequal costs for different types of misclassification.

By contrast, given a weighting function W on the outputs, the empirical loss function of ϕ given a training set can be computed as $\frac{1}{n}\sum_{i=1}^{n} W(y_i)I[y_i\phi(\mathbf{x}_i) < 0]$. The Bayes rule ϕ^* minimizes the expected loss function and is given by

$$\phi^*(\mathbf{x}) = \begin{cases} +1 & \text{if} \quad \dfrac{p(\mathbf{x})}{1 - p(\mathbf{x})} > \dfrac{c^+}{c^-} \\ -1 & \text{if} \quad \dfrac{p(\mathbf{x})}{1 - p(\mathbf{x})} < \dfrac{c^+}{c^-}. \end{cases} \tag{2}$$

It can be shown that $\phi^*(\mathbf{x}) = \mathrm{sign}[p(\mathbf{x}) - \frac{c^+}{c^+ + c^-}]$. By defining $W(-1) = c^+$ and $W(+1) = c^-$, we can express the Bayes rule as $\phi^*(\mathbf{x}) = \mathrm{sign}[p(\mathbf{x}) - \frac{W(-1)}{W(+1)+W(-1)}]$.

Recently, Qiao and Liu (2008) introduced a new criterion for classification performance called MWGE (*Mean Within Group Error*). This criterion considers the misclassification cost for each class individually, and the best classifier under this criterion should minimize the average of misclassification cost within each class. If two classes are extremely unbalanced, it can be more reasonable to use MWGE to ensure reasonable representation of the minority class. Once employing MWGE as the criterion, the loss function will be rewritten as $L[\phi] = \frac{c^+}{\pi^-} I[y = -1]I[\phi(\mathbf{x}) = +1] + \frac{c^-}{\pi^+} I[y = +1]I[\phi(\mathbf{x}) = -1]$. Given $\mathbf{x}$, the expected loss

is $E[L(\phi)|\mathbf{X} = \mathbf{x}] = \frac{c^+}{\pi^-}[1 - p(\mathbf{x})]I[\phi(\mathbf{x}) = +1] + \frac{c^-}{\pi^+}p(\mathbf{x})I[\phi(\mathbf{x}) = -1]$. Under the MWGE criterion, the corresponding Bayes rule ϕ_* is given by

$$\phi_*(\mathbf{x}) = \begin{cases} +1 & \text{if} \quad \dfrac{p(\mathbf{x})}{1-p(\mathbf{x})} > \dfrac{c^+\pi^+}{c^-\pi^-} \\ -1 & \text{if} \quad \dfrac{p(\mathbf{x})}{1-p(\mathbf{x})} < \dfrac{c^+\pi^+}{c^-\pi^-}. \end{cases}$$

2.1.2 Biased Sampling

In some real situations, the proportions in the sample may not reflect those in the target population due to sampling bias. For example, if the two classes have very different proportions in the population, the smaller class may be over-sampled, while the larger class may be under-sampled in order to achieve more balance in the sample.

Proportions	+1 class	−1 class
in population	π^+	π^-
in sample	π_s^+	π_s^-

Table 2: Proportions in the target population and the sample.

Assume the proportions are labelled as in Table 2. Let $(\mathbf{X}_s, Y_s)$ be a random pair that has the same distribution as the sample. Then the conditional probability of a subject $\mathbf{X}_s = \mathbf{x}$ in the sample belonging to the +1 class is

$$p_s(\mathbf{x}) = \mathrm{P}(Y_s = +1|\mathbf{X}_s = \mathbf{x}) = \frac{\pi_s^+ g^+(\mathbf{x})}{\pi_s^+ g^+(\mathbf{x}) + \pi_s^- g^-(\mathbf{x})}. \tag{3}$$

Comparing (1) and (3), the relationship of proportion ratios between the population and the sample is $\dfrac{p(\mathbf{x})}{1-p(\mathbf{x})} = \dfrac{\pi^+ g^+(\mathbf{x})}{\pi^- g^-(\mathbf{x})} = \dfrac{\pi_s^+ g^+(\mathbf{x})}{\pi_s^- g^-(\mathbf{x})}\dfrac{\pi^+\pi_s^-}{\pi^-\pi_s^+} = \dfrac{p_s(\mathbf{x})}{1-p_s(\mathbf{x})}\dfrac{\pi^+\pi_s^-}{\pi^-\pi_s^+}$. Then the Bayes rule in (2) can be expressed in terms of the π's by

$$\phi^*(\mathbf{x}) = \begin{cases} +1 & \text{if} \quad \dfrac{p_s(\mathbf{x})}{1-p_s(\mathbf{x})} > \dfrac{c^+\pi_s^+\pi^-}{c^-\pi_s^-\pi^+} \\ -1 & \text{if} \quad \dfrac{p_s(\mathbf{x})}{1-p_s(\mathbf{x})} < \dfrac{c^+\pi_s^+\pi^-}{c^-\pi_s^-\pi^+}. \end{cases}$$

If we define $W(+1) = c^{-}\pi_s^{-}\pi^{+}$ and $W(-1) = c^{+}\pi_s^{+}\pi^{-}$, the Bayes rule can be expressed as

$$\phi^*(\mathbf{x}) = \text{sign}[p_s(\mathbf{x}) - \frac{W(-1)}{W(+1) + W(-1)}]. \tag{4}$$

Now consider the situation when we use the MWGE criterion. The Bayes rule ϕ_* is then given by

$$\phi_*(\mathbf{x}) = \begin{cases} +1 & \text{if} \quad \dfrac{p_s(\mathbf{x})}{1 - p_s(\mathbf{x})} > \dfrac{c^{+}\pi_s^{+}}{c^{-}\pi_s^{-}} \\ -1 & \text{if} \quad \dfrac{p_s(\mathbf{x})}{1 - p_s(\mathbf{x})} < \dfrac{c^{+}\pi_s^{+}}{c^{-}\pi_s^{-}}. \end{cases}$$

Additional discussion of nonstandard situations for the SVM, including unequal cost and biased sampling, can be found in Lin, Lee and Wahba (2002).

2.2 Optimal DWD Weighting Scheme

Under the overall misclassification criterion in the situation of the previous subsection, for each $i = 1, \cdots, n$, we define the weight

$$W(y_i) = \begin{cases} c^{-}\pi_s^{-}\pi^{+} & \text{if} \quad y_i = +1 \\ c^{+}\pi_s^{+}\pi^{-} & \text{if} \quad y_i = -1. \end{cases} \tag{5}$$

The optimization task of the linear weighted DWD is to find a separating hyperplane with a d-dimensional normed vector ω and intercept b, which solve

$$\min_{\omega,b,\xi} \sum_{i=1}^{n} W(y_i)\Big(\frac{1}{r_i} + C\xi_i\Big), \quad \text{s.t.} \quad r_i = y_i(\mathbf{x}_i'\omega + b) + \xi_i, \ \frac{1}{2}\omega'\omega = \frac{1}{2}, \ r_i \geq 0, \ \xi_i \geq 0 \text{ for } i = 1, 2, \cdots, n. \tag{6}$$

Here we assign different weights to each data point according to the different costs and class proportions. Note that from the minimization problem (6) we do not necessarily need the exact values for $W(+1)$ and $W(-1)$. Instead it is enough to know the ratio of the two.

2.2.1 Equivalent Formulation

If for each $i = 1, ..., n$, define $f_i = f_i(\omega, b) = \mathbf{x}_i'\omega + b$, then the weighted DWD in (6) is equivalent to the optimization problem as follows.

$$\min_{\omega,b}\min_{\xi}\sum_{i=1}^{n}W(y_i)\Big(\frac{1}{y_i f_i+\xi_i}+C\xi_i\Big),\quad \text{s.t.}\quad \frac{1}{2}\omega'\omega=\frac{1}{2},\ y_i f_i+\xi_i\geq 0,\ \xi_i\geq 0\quad \text{for}\quad i=1,...,n. \tag{7}$$

It can be shown that the optimal solution for the inside optimization part of (7) is given by $\xi^*=(\xi_1^*,\cdots,\xi_n^*)$, where $\xi_i^*=\frac{1}{\sqrt{C}}-y_i f_i$ if $y_i f_i\leq\frac{1}{\sqrt{C}}$; $\xi_i^*=0$ otherwise. Then (7) amounts to

$$\min_{\omega,b}\sum_{i=1}^{n}W(y_i)\Big([2\sqrt{C}-C\cdot y_i f_i]I[y_i f_i\leq\frac{1}{\sqrt{C}}]+\frac{1}{y_i f_i}I[y_i f_i>\frac{1}{\sqrt{C}}]\Big),\quad \text{s.t.}\quad \frac{1}{2}\omega'\omega=\frac{1}{2}.$$

This representation provides some insights as a modification of the *Hinge loss* of the SVM.

2.2.2 Fisher Consistency

Define the *DWD loss function*

$$V(yf)=\begin{cases}2\sqrt{C}-C\cdot yf & \text{if}\quad yf\leq\frac{1}{\sqrt{C}}\\ \dfrac{1}{yf} & \text{otherwise.}\end{cases} \tag{8}$$

Then the weighted DWD optimization is $\min_{\omega,b}\sum_{i=1}^{n}W(y_i)V(y_i f_i(\omega,b))\quad \text{s.t.}\quad \frac{1}{2}\omega'\omega=\frac{1}{2}$. For any classification function f, the expected loss, i.e. the risk, is $R(f)=E[W(Y_s)V(Y_s f(\mathbf{X}_s))]=E[E\{W(Y_s)V(Y_s f(\mathbf{X}_s))|\mathbf{X}_s=\mathbf{x}_s\}]$.

Fisher consistency can be proved by showing that the sign of the global minimizer of unconditional risk is equal to the Bayes rule ϕ^* given in (4). Theorem 1 proves this relationship and thus Fisher consistency of the weighted DWD.

THEOREM 1 *Let f^* be the global minimizer of $E[W(Y_s)V(Y_s f(\mathbf{X}_s))]$, where V is given in (8). Then $sign[f^*(\mathbf{x})]=\phi^*(\mathbf{x})$, where $\phi^*(\mathbf{x})$ is given in (4), or equivalently, $sign[f^*(\mathbf{x})]=sign[p_s(\mathbf{x})-\frac{W(-1)}{W(+1)+W(-1)}]$.*

2.2.3 The MWGE criterion

Similarly, under the MWGE criterion, we can define the weight $W^*(y_i)=c^-\pi_s^-$, if $y_i=+1$; $c^+\pi_s^+$ if $y_i=-1$. DWD with this weighting scheme can also be shown to be Fisher consistent.

The proof is similar to that of Theorem 1 and is omitted here to save space. For simplicity, we define in general w_+ and w_- as the weights for either of the schemes $W(\cdot)$ or $W^*(\cdot)$.

3 HDLSS Asymptotics for Weighted DWD

In this section, we explore the HDLSS asymptotics of DWD. We first study the definition and representation of the two classes in the HDLSS data setting.

Datasets with more variables than observations tend to have surprising and counter-intuitive geometrical structures. Hall et al. (2005) took a d-asymptotics approach and showed that, under some conditions, pairwise distances between vectors are approximately constant, so that each data point in a sample of size n is located near a vertex of a regular n-simplex in $\mathbb{R}^d$. With two class populations in the binary classification problem, data points from the positive class $\mathcal{X}^+$ (size n^+) and those from the negative class $\mathcal{X}^-$ (size n^-) can be viewed as an n^+-simplex and an n^--simplex within the d-dimensional space respectively.

The results in Hall et al. (2005) require the entries of the data vector to be *nearly independent*, in the sense that when they are viewed as a time series with time index d, these entries must satisfy a $\rho-$mixing condition. Ahn et al. (2007) gave a milder condition for the result of Hall et al. (2005) to represent one HDLSS sample using asymptotic properties of the sample covariance matrix and its eigenvalues. However, a counterexample due to John Kent suggests that their result needs an additional condition.

In this section, we will first improve their theory by reconciling the assumption, then extend their work to geometrically represent two HDLSS data sets under mild conditions. Further, based on this representation, we explore the HDLSS asymptotics for DWD.

3.1 Geometric Representations for One Sample under Mild Conditions

First consider the positive class $\mathcal{X}^+(d) = \{\mathbf{x}_1^+(d), \mathbf{x}_2^+(d), \ldots, \mathbf{x}_{n^+}^+(d)\}$ with n^+ data vectors and d variables. We have a $d \times n^+$ data matrix $\mathbf{X}_d^+ = [\mathbf{x}_1^+, \mathbf{x}_2^+, \cdots, \mathbf{x}_{n^+}^+]$ with $d > n^+$, where $\mathbf{x}_j^+ = (x_{1j}^+, x_{2j}^+, \cdots, x_{dj}^+)^T \in \mathbb{R}^d$, $j = 1, 2, \cdots, n^+$, are independent and identically distributed from a

$d-$dimensional multivariate distribution with positive definite covariance matrix Σ_d^+. Without loss of generality, we assume that each $\mathbf{x}_j^+$ has zero mean. Denote the $d \times d$ sample covariance matrix of $\mathbf{X}_d^+$ as $S_d^+ = n_+^{-1}\mathbf{X}_d^+\mathbf{X}_d^{+T}$. The eigenvalue decomposition of Σ_d^+ is $\Sigma_d^+ = V_d^+\Lambda_d^+V_d^{+T}$, where $\Lambda_d^+ = diag\{\lambda_1^+, \cdots, \lambda_d^+\}$ is the eigenvalue diagonal matrix. Also we define the average of the eigenvalues $\sigma_d^2 = \frac{1}{d}\Sigma_{i=1}^d \lambda_{i,d}^+$. We can write $\mathbf{X}_d^+ = V_d^+\Lambda_d^{+1/2}\mathbf{Z}_d^+$, where $\mathbf{Z}_d^+ = \Lambda_d^{+-1/2}V_d^{+T}\mathbf{X}_d^+$ is a $d \times n^+$ random data matrix from a distribution with zero mean and identity covariance matrix. The $n^+ \times n^+$ dual sample covariance matrix is defined as $S_{D,d}^+ = d^{-1}\mathbf{X}_d^{+T}\mathbf{X}_d^+$. Denote the $n^+ \times n^+$ matrix $W_{i,d}^+$ as $(Z_{i,d}^+)^T Z_{i,d}^+$, where $Z_{i,d}^+$, $i = 1, 2, \cdots, d$, are the row vectors of $\mathbf{Z}_d^+$. It was noted in Ahn et al. (2007) that $dS_{D,d}^+$ has a simple Wishart representation,

$$dS_{D,d}^+ = \sum_{i=1}^d \lambda_{i,d}^+ W_{i,d}^+. \tag{9}$$

Note that if $\mathbf{X}_d^+$ is Gaussian, each $W_{i,d}^+$ follows independently the Wishart distribution $\mathcal{W}_{n^+}(1, I_{n^+})$.

Assumption 1 *For a fixed n^+, consider a sequence of random data matrices $\mathbf{X}_1^+, \cdots, \mathbf{X}_d^+, \cdots$, indexed by the number of rows d. Assume each $\mathbf{X}_d^+$ comes from a multivariate distribution with dimension d. Let $\lambda_{1,d}^+ \geq \cdots \geq \lambda_{d,d}^+$ be the eigenvalues of the covariance matrix Σ_d^+, and let $S_{D,d}^+$ be the corresponding $n^+ \times n^+$ dual sample covariance matrix. We assume the following,*

- *Each column of $\mathbf{X}_d^+$ has zero mean and covariance matrix Σ_d^+.*
- *The fourth moments of each entry of each column are uniformly bounded by $M^+ > 0$ and also the representation in (9) holds for each $\mathbf{X}_d^+$.*
- *Entries of $Z_d^+ = \Sigma_d^{+-\frac{1}{2}}\mathbf{X}_d^+ = \Lambda_d^{+-\frac{1}{2}}V_d^{+T}\mathbf{X}_d^+$ (as defined above) are independent.*
- *The eigenvalues of Σ_d^+ are sufficiently diffused, in the sense that*

$$\epsilon_d^+ = \frac{\Sigma_{i=1}^d (\lambda_{i,d}^+)^2}{(\Sigma_{i=1}^d \lambda_{i,d}^+)^2} \to 0 \ \ as \ \ d \to \infty. \tag{10}$$

- *The sum of the eigenvalues of Σ_d^+ is the same order as d, i.e. $\sigma_d^2 = O(1)$.*

Assumption 1 is a strengthening of the conditions of Ahn et al. (2007), in particular the third part did not appear there. Condition (10) can be viewed as a measure of the sphericity of the data matrix. Assumption 1 restricts the underlying distribution to be not too close to the extreme case of a few dominant eigenvalues. The spherical Gaussian is an example which has perfect sphericity, i.e. $\epsilon_d = \frac{1}{d}$. As mentioned in Ahn et al. (2007), the $\rho-$mixing condition in Hall et al. (2005) is also a special case that satisfies Assumption 1.

One main result of Ahn et al. (2007) is that under their weaker version of Assumption 1, the sample eigenvalues behave as if they follow an identity covariance matrix, in the sense that $\frac{1}{\sigma^2} S_{D,d} \rightarrow I^n, \quad as \quad d \rightarrow \infty$. Based on this theory they claim that the pairwise squared distance between the data vectors from $\mathcal{X}^+(d)$, rescaled by $\frac{1}{d}$, is approximately constant. However, John Kent pointed out that an additional assumption is needed, using a counterexample, in which $\mathbf{X}_d^+$ follows a mixed Normal distribution, i.e. with half probability $\mathbf{X}_d^+$ follows $N_d(0, I_d)$ and with half probability $\mathbf{X}_d^+$ follows $N_d(0, 10I_d)$. This example satisfies the moment conditions, eigenvalue sphericity condition and eigenvalue order condition. But the pairwise distances have a non-degenerate discrete limiting distribution.

The theory in Ahn et al. (2007) goes through if additional assumptions are added. A simple strengthening is to assume Gaussianity. Our Assumption 1 weakens this substantially by assuming only a set of underlying independent drivers, $\mathbf{Z}_d^+$. We restate the theorem as follows.

THEOREM 2 *Under Assumption 1, the dual sample covariance matrix, rescaled by σ_d^2, becomes approximately the identity matrix I_n, as $d \rightarrow \infty$.*

$$\frac{1}{\sigma_d^2} S_{D,d} \rightarrow I_n \ \ in \ \ probability, \ \ as \ \ d \rightarrow \infty.$$

A direct consequence of Theorem 2 is that the pairwise squared distance rescaled by d^{-1} is approximately constant as $d \rightarrow \infty$.

COROLLARY 1 *Under Assumption 1, the pairwise distances between the n^+ data vectors are approximately the same. In particular, scaled by $1/d\sigma_d^2$, the squared distance satisfies,*

$$\frac{1}{d\sigma_d^2}\|\mathbf{x}_k^+ - \mathbf{x}_l^+\|^2 \to 2, \quad \textit{in probability}, \quad \textit{as} \quad d \to \infty.$$

Thus these n^+ data vectors form a regular n^+-simplex in $\mathbb{R}^d$.

3.2 Geometric Representations for Two Samples

The n^--point sample $\mathcal{X}^-(d) = \{\mathbf{x}_1^-(d), \mathbf{x}_2^-(d), \ldots, \mathbf{x}_{n^-}^-(d)\}$ is defined in the same way. In particular, here the average of the eigenvalues is defined as $\tau_d^2 = \frac{1}{d}\Sigma_{i=1}^d \lambda_{i,d}^-$. When the eigenvalues for the negative class data matrix are sufficiently diffused, i.e. $\epsilon_d^- = \frac{\Sigma_{i=1}^d (\lambda_{i,d}^-)^2}{(\Sigma_{i=1}^d \lambda_{i,d}^-)^2} \to 0$ as $d \to \infty$, in the same manner, the pairwise squared distances between the n^- data vectors are approximately the same,

$$\frac{1}{d\tau_d^2}\|\mathbf{x}_k^- - \mathbf{x}_l^-\|^2 \to 2, \quad \text{as} \quad d \to \infty. \tag{11}$$

Now we generalize the means of the two classes so that $\mathbf{x}_k^+$ and $\mathbf{x}_k^-$ have arbitrary means. We assumed that the squared distance between the means, rescaling by $1/d$, is a constant μ^2,

$$\frac{1}{d}\|E(\mathbf{x}^+) - E(\mathbf{x}^-)\|^2 \to \mu^2. \tag{12}$$

The motivation of this setting is that it allows simple insight into the classification problem where we have two classes with different means. For convenience, we assume that the limiting average eigenvalues exist, $\sigma_d^2 \to \sigma^2$ and $\tau_d^2 \to \tau^2$ as $d \to \infty$.

THEOREM 3 *Assume two independent data samples $\mathcal{X}^+(d)$ and $\mathcal{X}^-(d)$ satisfy Assumption 1, and that $E(\mathbf{x}^+)$ and $E(\mathbf{x}^-)$ satisfy (12). Assume $\sigma_d^2 \to \sigma^2$ and $\tau_d^2 \to \tau^2$ as $d \to \infty$. Then the squared distance between a data vector in $\mathcal{X}^+(d)$ and a data vector in $\mathcal{X}^-(d)$, divided by d, converges in probability to $l^2 = (\sigma^2 + \tau^2 + \mu^2)$ as $d \to \infty$: for any $\varepsilon > 0$,*

$$Pr[|\frac{1}{d}\|\mathbf{x}_k^+ - \mathbf{x}_l^-\|^2 - l^2| \geq \varepsilon] \to 0, \quad \textit{as} \quad d \to \infty.$$

Theorem 3 says that, if each of the two samples is sufficiently spherical, as $d \to \infty$, the pairwise average distance between two data vectors from each sample is approximately constant. Theorem 3 gives the interclass distances in the d-limit , while Corollary 1 and (11) give the intraclass distances. From these results, one can organize the linear discrimination possibilities as follows.

1. If μ^2 is so large that $\sigma^2 + \tau^2 + \mu^2$ is significantly greater than $2\sigma^2$ and $2\tau^2$, then the two simplices are far from each other, and thus as discussed in Section 3.3 and Section 3.4, there is a natural separating hyperplane, that will give good classification, i.e. is generalizable.

2. If μ^2 is so small that $\sigma^2 + \tau^2 + \mu^2 < 2\max(\sigma^2, \tau^2)$, then it is much harder than above to classify by linear discrimination as shown in Section 3.3 and the generalization ability is also weakened as discussed in Section 3.4.

3.3 Asymptotic Properties of the DWD Intercept

Let us illustrate the properties of DWD in the HDLSS data settings. Let O^+ be the centroid of the n^+-simplex $\mathcal{X}^+(d)$ and O^- the centroid of the n^--simplex $\mathcal{X}^-(d)$. As noted in Hall, et al. (2005), the DWD hyperplane is orthogonal to the line O^+O^- joining the centroids. Let P be any point on the interval O^+O^-. We want to find the conditions where P becomes the DWD cut-off point.

In Figure 1, let α and β be the distance from P to the centroids. As noted in Hall, et al. (2005), P lies on the weighted DWD hyperplane only when

$$\frac{\alpha}{\beta} = \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}. \tag{13}$$

This determines the DWD hyperplane, which is orthogonal to the line O^+O^- and passes through the point P which satisfies condition (13). The larger $\frac{w_+ n^+}{w_- n^-}$ is, the closer the cut-off point P will be to O^-, and thus the more likely a new data point is to be classified to $\mathcal{X}^+$. Theorem 4 shows the conditions under which a future data point is always correctly classified or misclassified.

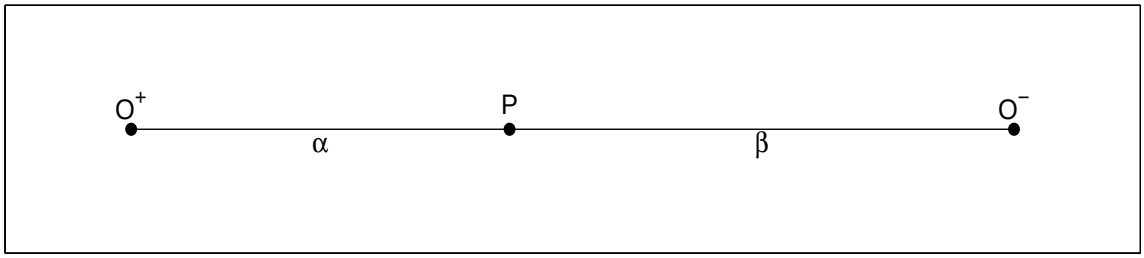

Figure 1: Simplex centroids O^+, O^- and the candidate DWD cut-off point P

THEOREM 4 *Assume that* $\sigma^2/[n_+^{\frac{3}{2}}w_+^{\frac{1}{2}}] \geq \tau^2/[n_-^{\frac{3}{2}}w_-^{\frac{1}{2}}]$*; if needed, interchange* X^+ *and* X^- *to satisfy this assumption.*

- *For a new data point* X_0^+ *from the* $\mathcal{X}^+-$*population,*

  1. *If* $\mu^2 > (n^-w_-/n^+w_+)^{\frac{1}{2}}\sigma^2/n^+ - \tau^2/n^-$*, then*

     $$Pr(X_0^+ \text{ is correctly classified by the weighted DWD}) \to 1, \text{ as } d \to \infty.$$

  2. *If* $\mu^2 < (n^-w_-/n^+w_+)^{\frac{1}{2}}\sigma^2/n^+ - \tau^2/n^-$*, then*

     $$Pr(X_0^+ \text{ is wrongly classified by the weighted DWD}) \to 1, \text{ as } d \to \infty.$$

- *For a new data point* X_0^- *from the* $\mathcal{X}^-$*–population, for any* $\mu > 0$*,*

  $$Pr(X_0^- \text{ is correctly classified by the weighted DWD}) \to 1, \text{ as } d \to \infty.$$

An intuitive interpretation of Theorem 4 is that the intraclass average variances σ^2 and τ^2, the sizes n^+ and n^- and the interclass squared distances μ^2, jointly control the ability to classify the new data point from $\mathcal{X}^+$ and $\mathcal{X}^+$. Large interclass distance will surely lead to better accuracy in general. When one class has smaller intraclass variance or larger sample size, traditional DWD will lead to a more accurate classification rule for it. This comes at a cost of worse classification performance for the other class. Weighted DWD helps to offset the effect of sample size to some extent.

Theorem 4 is the *weighted* version of Theorem 3 in Hall et al. (2005). Compared to its original version, Theorem 4 extends DWD by the introduction of w_+ and w_- into the conditions. For example, in the case of unbalanced data with equal cost and unbiased sampling, for relatively small n^- and large n^+, we have the weight ratio $\frac{w^+}{w^-} = \frac{n^-}{n^+}$. In Theorem 4, the main condition in Hall et al. (2005), $\sigma^2/n_+^{\frac{3}{2}} \geq \tau^2/n_-^{\frac{3}{2}}$, is relaxed to $\sigma^2/n^+ \geq \tau^2/n^-$. This condition is more easily satisfied so that, as shown in the theorem, one can classify a new data point from $\mathcal{X}^-$ correctly by the weighted DWD in contrast to the traditional DWD. However, the condition in Hall et al. (2005), under which the data point from $\mathcal{X}^+$ is correctly classified, $\mu^2 > (n^-/n^+)^{\frac{1}{2}}\sigma^2/n^+ - \tau^2/n^-$, becomes $\mu^2 > \sigma^2/n^+ - \tau^2/n^-$ now, which is not as easily attained as before.

To summarize, for the traditional DWD, misclassifications of some future points are unavoidable (in the asymptotic setting), because this is totally controlled by the relative magnitudes of μ^2, n^-, n^+, σ^2, τ^2, which are all aspects of the underlying distributions. However for the weighted DWD, we can adaptively choose the weights to adjust those relevant quantities, which will reduce the misclassified region and lead to better classification accuracy. In the ideal (but unrealistic) case, where the values μ^2, n^-, n^+, σ^2, τ^2 are known in advance, we can choose the weights intelligently such that the scenario (2) in Theorem 4 can be avoided to the extent possible, i.e., all the future points would be correctly classified.

3.4 Asymptotic Properties of the DWD Direction

Theorem 4 gives a sufficient condition under which new data are correctly classified with probability converging to 1. However, it holds under the assumption that the intraclass average variances σ^2 and τ^2, i.e. the noise levels, are not very large. Once the noise level, relative to the interclass distance μ^2, i.e. the signal level, is not ignorable, Theorem 4 cannot tell much about the performance of DWD, since the two samples are close and mixed. Instead, in this case, the similarity between the DWD direction (the vector orthogonal to the separating hyperplane) and the optimal linear classification direction (the line joining the

means of the two populations) is of more interest. If the angle between the above two directions is close to 0, the classification will be generalizable, in the sense of performing well for new data. Using the asymptotics of this angle, one can also evaluate the credibility of DWD as a classification rule.

THEOREM 5 *Consider $\mathcal{X}^+(d)$ and $\mathcal{X}^-(d)$ defined in Assumption 1. As $d \to \infty$, with probability converging to 1, the angle between the direction joining the means of two populations and the direction joining the centroids of n^+-simplex $\mathcal{X}^+(d)$ and n^--simplex $\mathcal{X}^-(d)$ becomes $\theta = cos^{-1}\left(\frac{\mu^2}{\mu^2+\sigma^2/n^+ +\tau^2/n^-}\right)^{\frac{1}{2}}$.*

Because the direction joins the data centroids coincides with the (standard or weighted) DWD direction asymptotically, the asymptotic property of the angle between the DWD direction and the optimal linear classification direction is implied by Theorem 5. In particular,

$$\theta \approx \begin{cases} 90^\circ, & \text{if } \mu^2 \ll \frac{\sigma^2}{n^+} + \frac{\tau^2}{n^-}, \\ 0^\circ, & \text{if } \frac{\sigma^2}{n^+} + \frac{\tau^2}{n^-} \ll \mu^2. \end{cases} \tag{14}$$

Consider intraclass average variances σ^2 and τ^2 as the noise level and the interclass distance μ^2 as the signal level. Theorem 5 and (14) imply that DWD tends to give the optimal linear classification direction when the signal is much greater than the noise, and also tends to give a direction which is orthogonal to the optimal one, when the noise is significantly greater than the signal. The second implication of Theorem 5 is that the angle goes to 0 if the sample sizes, n^+ and $n^- \to \infty$. This can be seen as a variation of Fisher Consistency in Theorem 1 from the $d-$asymptotic view of point.

4 Numerical Study

4.1 Simulation of 2-dimensional data

In this section, we consider a two-dimensional simulated example. The underlying population is unbalanced with fixed subpopulation proportions $\pi^+ = 10\%$ and $\pi^- = 90\%$. The sampling distribution is based on 400 samples with fixed proportions $\pi_s^+ = 40\%$ and $\pi_s^- = 60\%$.

For the purpose of tuning, we randomly divide each dataset into two halves: one is the training set and the other is the tuning set. Thus for both sets, $n^+ = 80$, $n^- = 120$. Various values of tuning parameter C in (6) are used and we choose the one minimizing the expected cost on the tuning set.

The positive subpopulation follows a bivariate normal distribution with mean $(0, 0)^T$ and covariance matrix I_2, whereas the negative subpopulation follows a bivariate normal with mean $(1.5, 1.5)^T$ and covariance matrix I_2. A testing set is generated from the population in the same way as the generation of the training and tuning sets, except that the proportion ratio for the testing set is set to be the same as that of the population, i.e. $\pi^+ = 10\%$, $\pi^- = 90\%$. The testing sample size is 600 ($n^+ = 60$, $n^- = 540$). Furthermore, we set the cost of a false negative to be twice of the cost of a false positive, i.e. $c^- = 2c^+$. We replicate the simulation 100 times.

4.1.1 Weighting Under the Overall Misclassification Criterion

According to (5), under the overall misclassification criterion, the weights for the data points are $W(+1) = 0.12$, $W(-1) = 0.36$. Under the same criterion, the Bayes rule ϕ^* is sign$[1.5 + \frac{2}{3}\log\frac{2}{9} - x_1 - x_2]$. The left panel of Figure 2 depicts the expected cost vs $\log_{10}(C)$ for the two versions of DWD as well as Bayes rule classifier on one typical simulation. In this sample, we choose $C = 10^1$ for the weighted DWD and $C = 10^{-\frac{1}{2}}$ for the standard DWD. We also tried the choice $C = 100/(dt)^2$ suggested in Marron, Todd and Ahn (2007) for the standard DWD, where dt is the median of the pairwise Euclidean distances between classes. Their choice gives $C = 13.254$, which is different for that of the standard DWD, but fairly close to the tuning parameter selected by the weighted DWD.

We compare the performance of the standard DWD (stdDWD), the weighted DWD (wDWD), and the Bayes rule (Bayes) on the training data on the left panel of Figure 3. It shows that the weighted DWD classifier lies closer to the Bayes rule than that of the standard DWD. Furthermore, both the weighted DWD and the Bayes rule hyperplanes are located far-

ther away from the positive class than the standard DWD. This is caused by the different cost ratio from that of the standard DWD. Table 3 gives the summary of classification results on the training set. Although the three methods have roughly the same number of misclassified points, the standard DWD produces the largest misclassification cost.

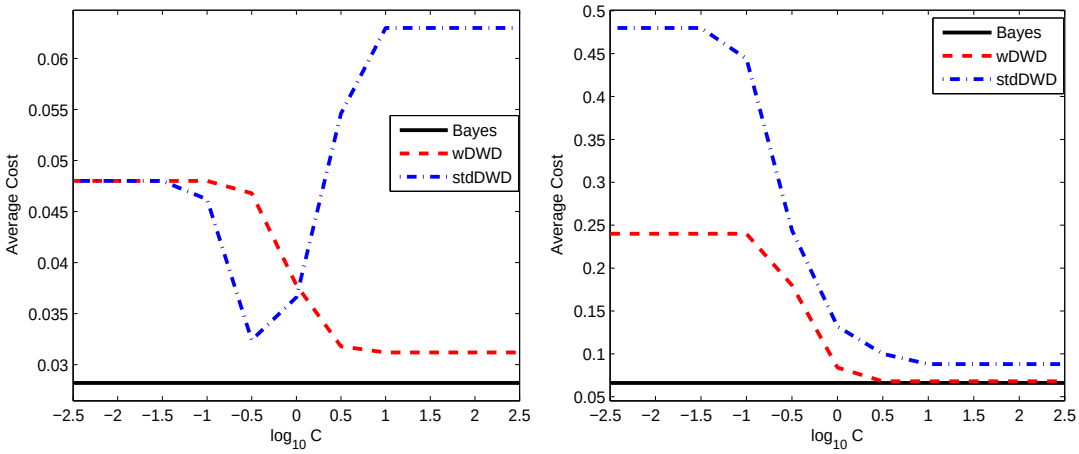

Figure 2: Parameter tuning for the Bayes classifier (Bayes), the weighted DWD (wDWD) and the standard DWD (stdDWD). The left panel displays the case of overall misclassification criterion. The right panel displays the case of MWGE criterion which will be discussed in the next section.

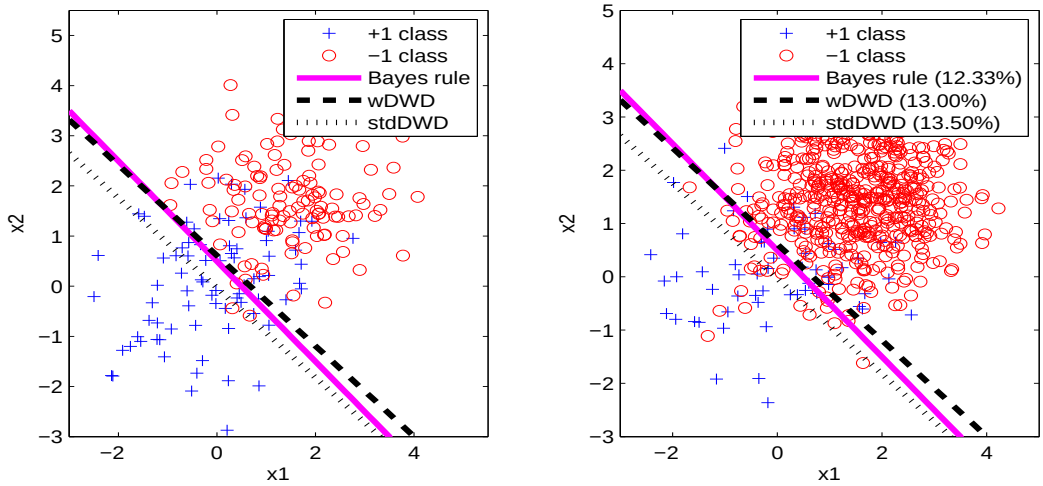

Figure 3: Classifications under the overall misclassification criterion for the training data (left panel) and the testing set (right panel): The percentages in the parentheses are the total misclassification rates.

To further compare the prediction performances of the weighted DWD with other methods, the right panel in Figure 3 shows the classification results on the testing data. We see that the weighted DWD and the Bayes rule tend to minimize the number of false negatives, which have a larger weight than false positives. However, standard DWD tends to balance the two types of misclassifications, since it treats the data points from the two classes equally without taking into account the different misclassification costs and the sampling bias. Table 4 summarizes the classification results on the testing set. The weighted DWD works better than the standard DWD both in term of the cost and total misclassification rate.

Three important observations from Figure 3, Table 3 and Table 4 are as follows:

1. The Bayes rule hyperplane and the weighted DWD hyperplane lie close to each other. As a consequence, the misclassification rates for the Bayes classifier and the weighted DWD are close. In contrast, the misclassification rate for the standard DWD is much worse.

Classifier	False negative(%)	False positive(%)	Total(%)	Total cost
Bayes(1)	36.10 (5.91)	3.74 (1.60)	16.69 (2.22)	2.54 (0.42)
wDWD(1)	34.97 (6.74)	3.60 (1.18)	16.15 (2.96)	2.46 (0.46)
stdDWD(1)	18.06 (3.80)	10.17 (2.29)	13.33 (2.33)	3.06 (0.57)

Table 3: The training errors: averaged misclassification rates for each class and the entire sample, as well as the total misclassification cost, over 100 replications. The numbers in the parentheses are standard errors.

Classifier	False negative(%)	False positive(%)	Total(%)	Total cost
Bayes(1)	37.37 (5.91)	3.91 (0.8)	7.26 (0.83)	10.99 (1.24)
wDWD(1)	40.67 (8.11)	3.53 (1.38)	7.24 (0.96)	11.31 (1.27)
stdDWD(1)	35.80 (13.37)	5.33 (3.13)	8.38 (1.81)	11.96 (1.49)

Table 4: The testing errors: averaged misclassification rates for each class and the entire sample, as well as the total misclassification cost, over 100 replications. The numbers in the parentheses are standard errors.

2. The misclassification rate of the weighted DWD for the testing data is similar to that for the training data, while the misclassification rate of the standard DWD changes dramatically when it is applied to the testing data. This is because the weighted DWD takes into account the biased sampling, i.e. it treats the training data as if it has the same proportion ratio as the testing data. Since the standard DWD does not incorporate the sampling bias, its generalization is worse than that of the weighted version.

3. Note that the false negative rates for both DWDs are much higher than the false positive rates. This is caused by the unbalanced population ($1 : 9$) as well as the use of the overall misclassification rate criterion (Qiao and Liu 2008). An alternative solution is to employ the MWGE and to use a flexible weighting scheme which is adjusted according to both nonstandard situation and the unbalanced data.

4.1.2 Weighting Under the MWGE criterion

In this section, we make use of the MWGE criterion (Qiao and Liu 2008) to calculate the new weighting scheme and the corresponding Bayes rule. Under the MWGE criterion, $W_*(+) = c^-\pi_s^-$, and $W_*(-) = c^+\pi_s^+$. The Bayes rule ϕ_* becomes $\text{sign}[1.5 + \frac{2}{3}\log 2 - x_1 - x_2]$. The right panel of Figure 2 shows the mean within group cost vs $\log_{10}(C)$ for the three classifiers. Figure 4 shows the classification plots on the training data and the testing data.

In Table 5, we include a column for the MWG cost(I) which displays the MWG cost from the classifiers under the overall misclassification criterion in the previous section. As expected, now the MWG cost(II) is much less compared to MWG cost(I). This is because with the new criterion MWGE, all three methods try to find an optimal classifier which minimizes the MWG cost. However, between the weighted DWD and the standard DWD, the weighted DWD improves even more. Moreover, on the testing set, the within group misclassification rates were balanced to 7.93% and 26.43% as compared to the ones in Table 4, 40.67% and 3.53% (error of the majority increased and error of the minority decreased).

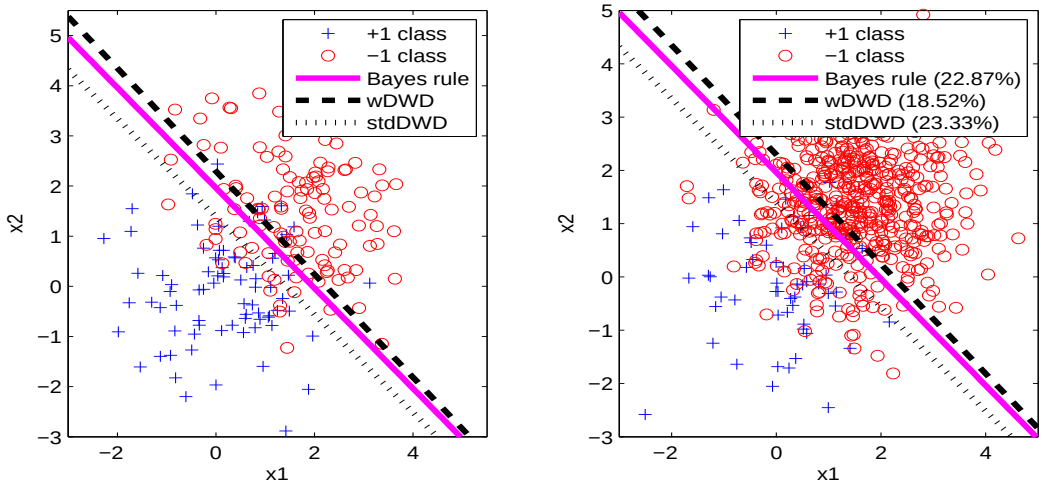

Figure 4: Classifications under the MWGE criterion on the training (left panel) and the testing: The percentages in the parentheses are the mean within group misclassification rates.

	Training Data		Testing Data			
Classifier	False negative	False positive	False negative	False positive	MWG cost	
(%)	$c^- = 2$	$c^+ = 1$	$c^- = 2$	$c^+ = 1$	(II)	(I)
Bayes(2)	8.59 (2.96)	22.72 (4.24)	8.77 (3.80)	23.42 (2.05)	20.48 (4.07)	39.53 (5.84)
wDWD(2)	8.20 (2.00)	22.12 (4.19)	7.93 (4.17)	26.43 (4.86)	21.15 (3.89)	42.43 (7.68)
stdDWD(2)	18.06 (3.80)	10.17 (2.29)	20.87 (5.80)	10.61 (2.03)	26.17 (5.43)	38.46 (12.06)

Table 5: Training and test error: Averaged misclassification rates for each class and the entire sample, as well as the mean within group cost (MWG cost), over 100 replications on both training and testing data. The fifth column labeled as (II) is the MWG cost corresponding to Bayes(2), wDWD(2) and stdDWD(2), while the sixth column labeled as (I) is the one for Bayes(1), wDWD(1) and stdDWD(1) from the previous section only for the purpose of comparison. The numbers in the parentheses are standard errors.

4.2 HDLSS Data

4.2.1 Simulated Data

Consider a typical HDLSS context. Let the dimension $d = 1000$, and the sample size of the training data $n = 200$. Assume the data is balanced with $\pi^+ = \pi^- = 50\%$ and equal cost

$c^- = c^+$, but with biased sampling, $\pi_s^+ = 20\%$ and $\pi_s^- = 80\%$. The weights for this data set are $w_+ = 0.4$, and $w_- = 0.1$. Note that because $\pi^+ = \pi^-$, the two weighting schemes under both criteria are the same.

The positive data vectors follow a 1000-dimensional normal distribution $X^+ \sim N(u\mathbf{1}_d, \mathbf{I}_d)$, where $\mathbf{1}_d = [1, 1, \cdots, 1]^T$, and the negative data vectors follow $X^- \sim N(-u\mathbf{1}_d, \mathbf{I}_d)$. Here we let $u = \frac{1}{2}\sqrt{5^2/d}$ so that the distance between the means of the two populations is 5. For tuning and testing purposes, we generate the tuning set of size 200 and testing set of size 600. We replicate the simulation 100 times.

Figure 5 shows the misclassification error for one realization of tuning data set, over different tuning parameter values C, for the Bayes rule, the standard DWD and the weighted DWD.

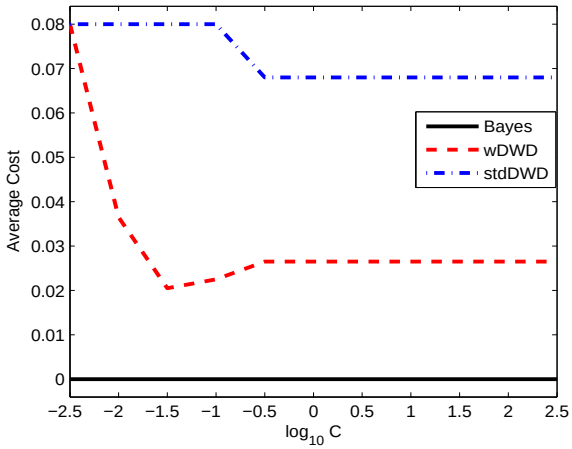

Figure 5: Parameter tuning for HDLSS example.

The parameter tuning procedure gives the optimal value of C for the weighted DWD, $C = 10^{-1.5}$. The parameter specification $C = 100/(dt)^2$ suggested in Marron, Todd and Ahn (2007) turns out to be $C = 0.0496 = 10^{-1.3046}$. However, the optimal value of C for the standard DWD is $C = 10^{-0.5}$. This implies that the suggestion by Marron, Todd and Ahn (2007) serves as a good parameter specification for the weighted DWD in the HDLSS case, but not for the standard DWD in the biased sampling case.

In Figure 6, we project all the training and testing data onto the weighted DWD direction

and the standard DWD direction respectively. The DWD separating hyperplane intersects this direction at the dashed vertical line. One can see that although both classifiers give perfect classification of the training dataset, the weighted DWD outperforms the standard DWD on the testing data, and thus has better generalization.

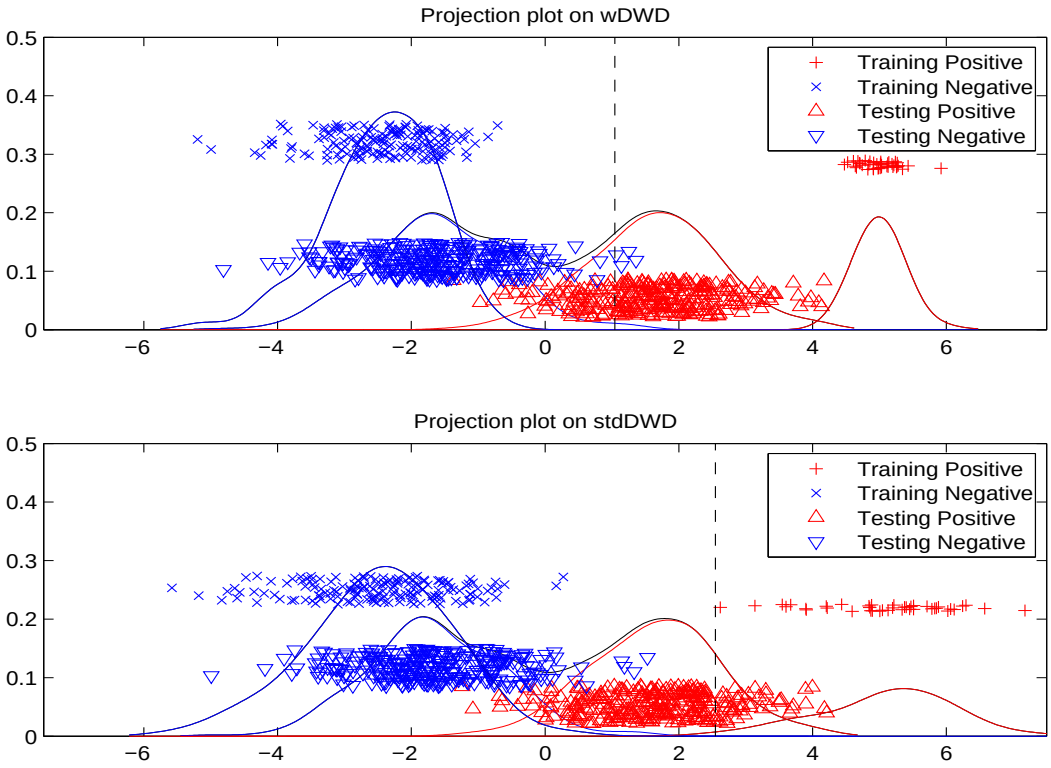

Figure 6: Projection plot of data points onto wDWD and stdDWD directions: 'Plus' represents positive training data set points. 'Cross' represents negative training data. 'Upward triangle' represents positive testing data. 'Downward triangle' represents negative testing data. The DWD separating hyperplanes intersects the wDWD and stdDWD directions at the two dashed vertical lines respectively. This shows much better performance of weighted DWD in this HDLSS biased sampling situation.

The summary table for training is omitted, since both methods work perfectly on the training set. Table 6 summarizes the misclassification on the testing dataset. Weighted DWD works much better than standard DWD in this high-dimensional data setting, in contrast to the slight improvement of the weighted DWD in the low-dimensional case in the previous

section.

Classifier	False negative(%)	False positive(%)	Total cost / MWG cost
Bayes(1, 2)	0.56 (0.41)	0.62 (0.46)	0.59 (0.30)
wDWD(1, 2)	25.05 (10.04)	3.09 (9.30)	14.07 (3.57)
stdDWD(1, 2)	80.37 (3.61)	0.003 (0.03)	40.19 (1.81)

Table 6: The testing errors: misclassification rates for each class, as well as the total misclassification cost (or MWG cost in this case), over 100 replications. The numbers in the parentheses show standard error.

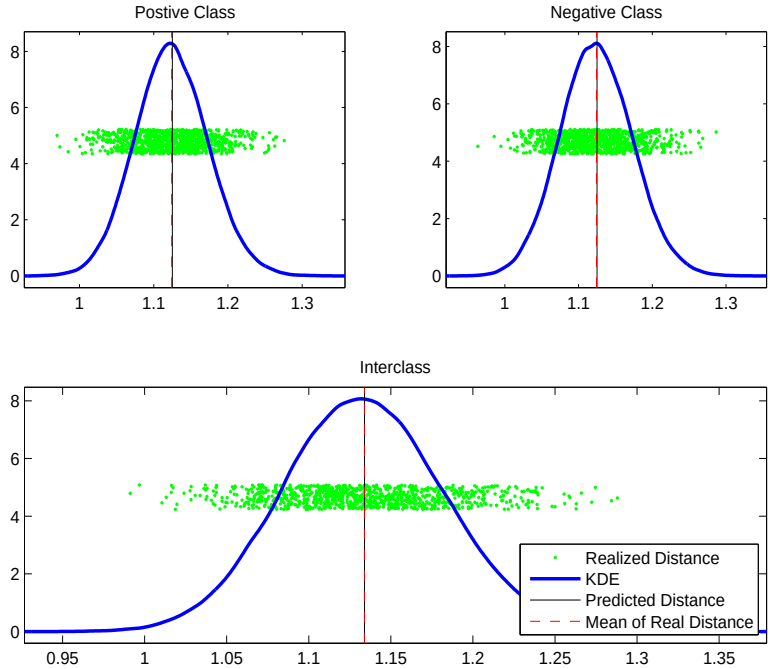

Figure 7: Kernel density estimation of the rescaled squared intraclass distances for both classes, and the interclass pairwise distances. Observed distances are shown as green dots.

In order to verify Theorem 5, we calculate the angle between the weighted DWD direction and the optimal linear classification direction $[1, 1, \cdots, 1]^T$ over the 100 simulated replications. In this simulation, the squared interclass distance rescaled by d^{-1} is $\mu^2 = (2u)^2 = 25/1000$. The interclass average variances for both classes, σ^2 and τ^2, are taken to be 1. The sample sizes are $n^+ = 40$ and $n^- = 160$. Thus according to Theorem 5, as

$d \to \infty$ (here $d = 1000$), the angle would become approximately $\theta = cos^{-1}\left(\frac{\mu^2}{\mu^2+\sigma^2/n^+ +\tau^2/n^-}\right)^{\frac{1}{2}} = cos^{-1}(\sqrt{0.4444}) = 48.19°$. The angles reported in the simulation study give a mean of 48.97° and standard error of 1.21°. Thus our theoretical angle falls into the one standard deviation range. This relatively large angle, which implies large deviation of DWD from the optimal linear classification direction, is expected because of the relatively large noise level.

In order to compare the theoretical interclass and intraclass distances with the realized ones, we calculate and report the pairwise squared distances rescaled by d^{-1} for the positive class, the negative class and interclass respectively. Figure 7 depicts the kernel density estimate for the rescaled squared distances and Table 7 summarizes the statistics. From Table 7, all of the mean rescaled squared distances fall reasonably close to the theoretical prediction, according to Theorem 2 and Theorem 3.

statistics	Positive	Negative	Interclass
number of pairs	72010	191890	235600
mean	2.0038	1.9992	2.0265
(theoretical)	$(2\sigma^2 = 2)$	$(2\tau^2 = 2)$	$(\sigma^2 + \tau^2 + \mu^2 = 2.025)$
s.d.	0.0896	0.0882	0.0903
skewness	0.1158	0.0936	0.1004
kurtosis	3.0545	3.0272	3.0466

Table 7: Summary statistics for the rescaled pairwise squared distances.

4.2.2 Real Data

In this section we show the performance of the weighted DWD and its asymptotics by running both the weighted and the standard DWD's on the Human Lung Carcinomas Microarrays Dataset (available from http://www.broad.mit.edu/mpr/lung/). In this dataset, there are six classes: 'Adenocarcinoma', 'Squamous', 'Pulmonary', 'Colon', 'Normal' and 'Small Cell Carcinoma', with sample sizes 128, 21, 20, 13, 17 and 6 respectively. We combine the

last four and the first two subclasses to form the positive and negative classes respectively. These samples have been described in detail previously (Bhattacherjee et al. 2001; Meyerson and Hayes 2005) and are analyzed by Liu et al. (2008). After some initial gene filtering, the dataset contains 2530 genes.

In each experiment, we selected 140 cases for the training set with 40 positive cases and 100 negative cases. The rest of the samples form the testing set with 16 being positive and 49 being negative. Within the 140 training data set, we split it into 5 groups and make 5-fold cross validation for tuning parameter selection. We keep the original proportion of each subclass in each block of the training set and the testing set so that each subclass has equal opportunity to appear in each block. The division between training and testing data is randomly repeated 100 times.

Assume that the cost is equal, i.e. $c^+ = c^- = 1$. Since the unbalanced data problem is too large to ignore, we use the weighting scheme $W_*(\cdot)$ under the MWGE criterion, i.e. $w_+ = c^-\pi_s^- = 100/140$ and $w_- = c^+\pi_s^+ = 40/140$.

Classifier	False negative(%)	False positive(%)	Total cost	MWG cost
wDWD(2)	3.31 (4.48)	4.78 (2.65)	4.42 (2.34)	4.04 (2.68)
stdDWD(2)	14.44 (8.41)	4.27 (2.53)	6.77 (2.19)	9.35 (3.88)

Table 8: The testing errors: averaged misclassification rates for each class and total sample, as well as the MWG cost, over 100 replications. The numbers in the parentheses show standard error.

Table 8 summarizes the performance of the weighted DWD and the standard DWD on the testing data set. At the cost of increasing the misclassification rate for the majority class by 4.78-4.27=0.51%, the weighted DWD improves the classification on the minority by 14.44-3.31=11.13%, improves the total misclassification rate by 6.77-4.42=2.35% and improves the MWG misclassification cost by 9.35-4.04=5.31%.

Estimator	$\hat{\sigma}^2$	$2\times\hat{\sigma}^2$	$\hat{\tau}^2$	$2\times\hat{\tau}^2$	$\hat{\mu}^2$	$\hat{\sigma}^2+\hat{\tau}^2+\hat{\mu}^2$	$\hat{\theta}$
Value	1.1071	2.2142	0.8563	1.7126	0.3541	2.3175	15.0268°

Table 9: Estimated $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{\mu}^2$, the former two multiplied by 2, the sum of all three, and the estimated angle.

In order to verify Theorem 4, we calculate the rescaled (by d^{-1}) average of estimated eigenvalues, $\hat{\sigma}^2$ for the positive class and $\hat{\tau}^2$ for the negative class respectively, as well as the rescaled squared distance between estimated means for each class $\hat{\mu}^2$. Table 9 shows the estimated $\hat{\sigma}^2$, $\hat{\tau}^2$ and $\hat{\mu}^2$ based on the whole data set. The estimated angle $\hat{\theta}$ between DWD direction and optimal linear classification direction is also given in Table 9.

	Positive		Negative	
Condition 1:	$\hat{\sigma}^2$	1.1071	$\hat{\tau}^2$	0.8563
	n_+	56	n_-	149
	w_+	100/140	w_-	40/140
	$\hat{\sigma}^2/[n_+^{\frac{3}{2}}w_+^{\frac{1}{2}}]$	**0.0031**	$\hat{\tau}^2/[n_-^{\frac{3}{2}}w_-^{\frac{1}{2}}]$	**0.0009**
Condition 2:	$\hat{\mu}^2$	$(n^-w_-/n^+w_+)^{\frac{1}{2}}\hat{\sigma}^2/n^+ - \hat{\tau}^2/n^-$		
	0.3541	**0.0146**		

Table 10: Verification of Theorem 4.

The estimated angle between the DWD direction and optimal linear classification direction is 15.0268°, which is somewhat better than the simulated data in the previous section. The reason is that here the signal level is relative greater than that in the simulated case. This leads to better DWD classifier in terms of angle. Moreover, the good performance of the weighted DWD can be explained by Theorem 4. It is also shown in Table 10 that the conditions for Theorem 4, $\sigma^2/[n_+^{\frac{3}{2}}w_+^{\frac{1}{2}}] \geq \tau^2/[n_-^{\frac{3}{2}}w_-^{\frac{1}{2}}]$ and $\mu^2 > (n^-w_-/n^+w_+)^{\frac{1}{2}}\sigma^2/n^+ - \tau^2/n^-$, are both satisfied, thus as long as the estimations of σ^2, τ^2 and μ^2 are correct , the weighted DWD can always classify new data point correctly.

5 Conclusion

In the article, we have proposed the weighted DWD for unbalanced datasets with possible unequal costs and biased sampling. We have made the following contributions. First of all, we have provided the optimal weighting schemes for several nonstandard situations. Secondly, we show Fisher Consistency for DWD. Thirdly, we represent data sets from two classes geometrically in HDLSS settings. Finally, we develop the HDLSS asymptotic properties of the (weighted) DWD. Both *n*-asymptotics and *d*-asymptotics are considered. Our numerical examples demonstrate the effectiveness of the weighted DWD and verify the asymptotic results.

The results of the tuning procedure in our simulated examples suggest that the choice of tuning parameter $C = 100/(dt)^2$ proposed by Marron, Todd and Ahn (2007), which was originally designed for balanced and standard data, also works well in unbalanced and nonstandard data cases as long as we use weighted DWD instead of standard DWD. We recommend the use of their simple recommendation for weighted DWD so that the readers can avoid the computational burden caused by trying many possible values of the parameter C.

Appendix

Proof of Theorem 1:

For any fixed $\mathbf{x}$, the conditional risk is

$$E[W(Y_s)V(Y_s f)|\mathbf{X}_s = \mathbf{x}] = p_s(\mathbf{x})W(+1)V(f(\mathbf{x})) + (1 - p_s(\mathbf{x}))W(-1)V(-f(\mathbf{x})).$$

For simplicity, we write $R(f) = p_s W(+1)V(f)+(1-p_s)W(-1)V(-f)$. Then f^* is obtained by solving $R'(f) = 0$, where $R'(f) = p_s W(+1)1V'(f) - (1 - p_s)W(-1)V'(-f)$. Straightforward computations give

$$V(f) = \begin{cases} 2\sqrt{C} - Cf & \text{if} \quad f \le \frac{1}{\sqrt{C}} \\ \dfrac{1}{f} & \text{otherwise,} \end{cases} \text{ and } V(-f) = \begin{cases} 2\sqrt{C} + Cf & \text{if} \quad f \ge -\frac{1}{\sqrt{C}} \\ -\dfrac{1}{f} & \text{otherwise.} \end{cases}$$

We can show that, for fixed p_s, $R(f)$ is continuous and differentiable everywhere and $R(f)$ is convex in $[-\infty, \infty]$, i.e. $R'(f)$ is nondecreasing. By solving the equation $R'(f) = 0$, we get the critical point f^*. Direct calculation gives us the minimizer of $R(f)$ as

$$f^* = \frac{1}{\sqrt{C}} \cdot \begin{cases} \sqrt{\frac{p_s W(+1)}{(1-p_s)W(-1)}} & \text{if} \quad \frac{p_s W(+1)}{(1-p_s)W(-1)} > 1 \\ 0 & \text{if} \quad \frac{p_s W(+1)}{(1-p_s)W(-1)} = 1 \\ -\sqrt{\frac{(1-p_s)W(-1)}{p_s W(+1)}} & \text{if} \quad \frac{p_s W(+1)}{(1-p_s)W(-1)} < 1. \end{cases}$$

Note when $\frac{p_s W(+1)}{(1-p_s)W(-1)} = 1$, f^* can take any value in $[-\sqrt{\frac{(1-p_s)W(-1)}{C p_s W(+1)}}, \sqrt{\frac{p_s W(+1)}{C(1-p_s)W(-1)}}]$. We choose 0 here for convenience. Therefore, the minimizer of $R(f)$ satisfies $\text{sign}[f^*] = \text{sign}[\frac{p_s W(+1)}{(1-p_s)W(-1)} - 1] = \text{sign}[p_s W(+1) - (1-p_s)W(-1)] = \text{sign}[p_s\{W(+1) + W(-1)\} - W(-1)] = \text{sign}[p_s > \frac{W(-1)}{W(+1)+W(-1)}] = \phi^*$.

Proof of Theorem 2:

Let $\mathbf{Z}_d^+ = \Lambda_d^{+-1/2} V_d^{+T} \mathbf{X}_d^+ = [\mathbf{z}_1^+, \cdots, \mathbf{z}_{n^+}^+]$, where $\mathbf{z}_k^+ = [z_{1k}^+, \cdots, z_{dk}^+]^T$ is the k-th column. Each column of $\mathbf{Z}_d^+$ is independently and identically distributed as an underlying d-dimensional distribution with identity covariance matrix I_d, where Λ_d^+ and V_d^+ form the eigenvalue-decomposition of the covariance matrix, $\Sigma_d^+ = V_d^+ \Lambda_d^+ V_d^{+T}$. Define the relative eigenvalue by $\tilde{\lambda}_{i,d}^+ = \lambda_{i,d}^+ / (\Sigma_{i=1}^d \lambda_{i,d}^+)$. The sphericity condition in Assumption 1 is equivalent to $\Sigma_{i=1}^d (\tilde{\lambda}_{i,d}^+)^2 \to 0$, as $d \to \infty$. Note that relative eigenvalues sum up to 1, i.e. $\Sigma_{i=1}^d \tilde{\lambda}_{i,d}^+ = 1$.

From the representation in (9), $\frac{1}{\sigma_d^2} S_{D,d}^+ = \frac{1}{\Sigma_{i=1}^d \lambda_{i,d}^+} \Sigma_{i=1}^d (\lambda_{i,d}^+ W_{i,d}^+) = \Sigma_{i=1}^d (\tilde{\lambda}_{i,d}^+ W_{i,d}^+)$. The $k-$th diagonal element of $\frac{1}{\sigma^2} S_{D,d}^+$ can be expressed as $\Sigma_{i=1}^d \tilde{\lambda}_{i,d}^+ (z_{ik}^+)^2$, where the z_{ik}^+'s $(i = 1, \cdots, d)$ are independent distributed with mean 0 and unit variance. And the (k,l)-th off-diagonal element of $\frac{1}{\sigma_d^2} S_{D,d}^+$ can be expressed as $\Sigma_{i=1}^d \tilde{\lambda}_{i,d}^+ (z_{ik}^+ z_{il}^+)$, where all z_{ik}^+'s and z_{il}^+'s are independent $(i = 1, \cdots, d)$, with mean 0 and unit variance.

Firstly, for diagonal element, $E[\Sigma_{i=1}^d \tilde{\lambda}_{i,d}^+ (z_{ik}^+)^2] = \Sigma_{i=1}^d \tilde{\lambda}_{i,d}^+ = 1$. By Chebyshev's inequality, for any $\varepsilon > 0$ and the uniform fourth moment bound M for z_i's, $Pr(|\Sigma_{i=1}^d \tilde{\lambda}_{i,d}^+ (z_{ik}^+)^2 - 1| >$

$\varepsilon) \leq \varepsilon^{-2} var(\Sigma_{i=1}^{d} \tilde{\lambda}_{i,d}^{+}(z_{ik}^{+})^2) \leq \varepsilon^{-2} M \Sigma_{i=1}^{d} (\tilde{\lambda}_{i,d}^{+})^2 \to 0, \ \ as \ \ d \to \infty$. Secondly, for off-diagonal element, $E[\Sigma_{i=1}^{d} \tilde{\lambda}_{i,d}^{+}(z_{ik}^{+} z_{il}^{+})] = 0$. By Chebyshev's inequality, for any $\varepsilon > 0$, $Pr(|\Sigma_{i=1}^{d} \tilde{\lambda}_{i,d}^{+}(z_{ik}^{+} z_{il}^{+}) - 0| > \varepsilon) \leq \varepsilon^{-2} var(\Sigma_{i=1}^{d} \tilde{\lambda}_{i,d}^{+}(z_{ik}^{+} z_{il}^{+})) = \varepsilon^{-2} \Sigma_{i=1}^{d} (\tilde{\lambda}_{i,d}^{+})^2 \to 0, \ as \ d \to \infty$. To summarize the analysis above, each element of $\frac{1}{\sigma_d^2} S_{D,d}^{+}$ converges to the counterpart of the identity matrix I_n in probability as $d \to \infty$.

Note that when each column of $\mathbf{X}_d^{+}$ follows multivariate Gaussian distribution, so is $\mathbf{z}_k^{+}$, the k-th column of Z_d^{+}. Hence, with identity covariance matrix of $\mathbf{z}_k^{+}$, its entries, $z_{ik}^{+} (i = 1 \cdots d)$, are independent, which satisfies the independence condition.

Proof of Corollary 1

Let $\mathbf{x}_j^{+} = (\mathbf{x}_{1j}^{+}, \cdots, \mathbf{x}_{dj}^{+})^T, \ j = 1, \cdots, n^{+}$, be the jth column of the data matrix $\mathbf{X}^{+}$. Let $\mathbf{x}_j^{-} = (\mathbf{x}_{1j}^{-}, \cdots, \mathbf{x}_{dj}^{-})^T, \ j = 1, \cdots, n^{-}$, be the jth column of the data matrix $\mathbf{X}^{-}$. The squared distance between $\mathbf{x}_k^{+}$ and $\mathbf{x}_l^{+}$, rescaled by $(d\sigma_d^2)^{-1}$ is $\frac{1}{d\sigma_d^2} \|\mathbf{x}_k^{+} - \mathbf{x}_l^{+}\|^2 = \frac{1}{d\sigma_d^2} \Sigma_{i=1}^{d} (\mathbf{x}_{ik}^{+} - \mathbf{x}_{il}^{+})^2 = \frac{1}{d\sigma_d^2} \Sigma_{i=1}^{d} (\mathbf{x}_{ik}^{+})^2 + \frac{1}{d\sigma_d^2} \Sigma_{i=1}^{d} (\mathbf{x}_{il}^{+})^2 - \frac{2}{d\sigma_d^2} \Sigma_{i=1}^{d} \mathbf{x}_{ik} \mathbf{x}_{il}$. The first and second terms on the right hand side are the k-th and l-th diagonal elements of $\frac{1}{\sigma_d^2} S_{D,d}^{+}$ respectively, which were proved to converge to 1 in probability as $d \to \infty$ in Theorem 2. The third term is the (k, l)-th off-diagonal element of $\frac{1}{\sigma_d^2} S_{D,d}^{+}$, which converges to 0 in probability as $d \to \infty$. Thus $\frac{1}{d\sigma_d^2} \|\mathbf{x}_k^{+} - \mathbf{x}_l^{+}\| \to 2$, in probability as $d \to \infty$.

LEMMA 1 *Assume that* $\Sigma_{i=1}^{d} (\lambda_{i,d}^{+})^2, \Sigma_{i=1}^{d} (\lambda_{i,d}^{-})^2 \to 0$, *as* $d \to \infty$ *and that* $\Sigma_{i=1}^{d} \lambda_{i,d}^{+} = \Sigma_{j=1}^{d} \lambda_{j,d}^{-} = 1$. *Denote* $U = [u_{ij}]_{i,j=1,\cdots,d}$ *as an arbitrary* $d \times d$ *orthogonal matrix. Then it holds that* $\Sigma_{i=1}^{d} \Sigma_{j=1}^{d} u_{i,j}^2 \lambda_{i,d}^{+} \lambda_{j,d}^{-} \to 0, \ \ as \ \ d \to \infty$.

Proof of Lemma 1: Note that sum of squared entries in each column and row of U is 1. By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\Sigma_{i=1}^{d} \Sigma_{j=1}^{d} u_{i,j}^2 \lambda_{i,d}^{+} \lambda_{j,d}^{-} = \ & \Sigma_{i=1}^{d} \lambda_{i,d}^{+} (\Sigma_{j=1}^{d} u_{i,j}^2 \lambda_{j,d}^{-}) && \leq \Sigma_{i=1}^{d} \lambda_{i,d}^{+} [\Sigma_{j=1}^{d} (\lambda_{j,d}^{-})^2]^{\frac{1}{2}} (\Sigma_{j=1}^{d} u_{i,j}^4)^{\frac{1}{2}} \\
\leq \ & \Sigma_{i=1}^{d} \lambda_{i,d}^{+} [\Sigma_{j=1}^{d} (\lambda_{j,d}^{-})^2]^{\frac{1}{2}} (\Sigma_{j=1}^{d} u_{i,j}^2)^{\frac{1}{2}} && = \Sigma_{i=1}^{d} \lambda_{i,d}^{+} [\Sigma_{j=1}^{d} (\lambda_{j,d}^{-})^2]^{\frac{1}{2}} \\
= \ & [\Sigma_{j=1}^{d} (\lambda_{j,d}^{-})^2]^{\frac{1}{2}} \Sigma_{i=1}^{d} \lambda_{i,d}^{+} && = [\Sigma_{j=1}^{d} (\lambda_{j,d}^{-})^2]^{\frac{1}{2}} \to 0, \ \text{as} \ d \to \infty.
\end{aligned}
$$

Proof of Theorem 3

Let $\mathbf{x}_j^+ = (\mathbf{x}_{1j}^+, \cdots, \mathbf{x}_{dj}^+)^T$, $j = 1, \cdots, n^+$, be the jth column of the data matrix $\mathbf{X}^+$. Let $\mathbf{x}_j^- = (\mathbf{x}_{1j}^-, \cdots, \mathbf{x}_{dj}^-)^T$, $j = 1, \cdots, n^-$, be the jth column of the data matrix $\mathbf{X}^-$. The squared distance between $\mathbf{x}_k^+$ and $\mathbf{x}_l^-$ is

$$\begin{aligned} \|\mathbf{x}_k^+ - \mathbf{x}_l^-\|^2 &= \Sigma_{i=1}^d \{[\mathbf{x}_{ik}^+ - E(\mathbf{x}_{i\cdot}^+)] - [\mathbf{x}_{il}^- - E(\mathbf{x}_{i\cdot}^-)] + [E(\mathbf{x}_{i\cdot}^+) - E(\mathbf{x}_{i\cdot}^-)]\}^2 \quad (15) \\ &= \Sigma_{i=1}^d (\dot{\mathbf{x}}_{ik}^+)^2 + \Sigma_{i=1}^d (\dot{\mathbf{x}}_{il}^-)^2 - 2\Sigma_{i=1}^d (\dot{\mathbf{x}}_{ik}^+)(\dot{\mathbf{x}}_{il}^-) \quad (16) \\ &\quad + \Sigma_{i=1}^d [E(\mathbf{x}_{i\cdot}^+) - E(\mathbf{x}_{i\cdot}^-)]^2 + 2\Sigma_{i=1}^d [E(\mathbf{x}_{i\cdot}^+) - E(\mathbf{x}_{i\cdot}^-)][\dot{\mathbf{x}}_{ik}^+ - \dot{\mathbf{x}}_{il}^-]. \quad (17) \end{aligned}$$

Here $\dot{\mathbf{x}}_{ik}^+ = \mathbf{x}_{ik}^+ - E(\mathbf{x}_{i\cdot}^+)$ and $\dot{\mathbf{x}}_{il}^- = \mathbf{x}_{il}^- - E(\mathbf{x}_{i\cdot}^-)$ are the ith entries on the kth and lth columns of the *de-meaned* data matrices $\dot{\mathbf{X}}^+$ and $\dot{\mathbf{X}}^-$.

The first two terms in (16), rescaled by $(d\sigma_d^2)^{-1}$ and $(d\tau_d^2)^{-1}$ respectively, are the kth and lth diagonal entries of $\frac{1}{\sigma_d^2} S_D^+$ and $\frac{1}{\tau_d^2} S_D^-$. By the proof of Theorem 2, both converge to 1 in probability as $d \to \infty$. Thus, for any $\varepsilon > 0$, $Pr(|\frac{1}{d}\Sigma_{i=1}^d (\dot{\mathbf{x}}_{ik}^+)^2 - \sigma^2| \geq \varepsilon) \to 0$, as $d \to \infty$ and $Pr(|\frac{1}{d}\Sigma_{i=1}^d (\dot{\mathbf{x}}_{il}^-)^2 - \tau^2| \geq \varepsilon) \to 0$, as $d \to \infty$.

The third term, $\Sigma_{i=1}^d (\dot{\mathbf{x}}_{ik}^+)(\dot{\mathbf{x}}_{il}^-)$, is the inner product of $\dot{\mathbf{x}}_k^+$ and $\dot{\mathbf{x}}_l^-$, the $k-$th column of the demeaned data matrix $\dot{\mathbf{X}}^+$, and the $l-$th column of the demeaned data matrix $\dot{\mathbf{X}}^-$. Recall that we can write $\dot{\mathbf{x}}_k^+ = V^+ \Lambda^{+1/2} \mathbf{z}_k^+$, where $\mathbf{z}_k^+ = (z_1^+, \cdots, z_d^+)^T$ is a d dimensional vector from a distribution with the identity covariance matrix and zero mean. So is $\dot{\mathbf{x}}_l^- = V^- (\Lambda^-)^{1/2} \mathbf{z}_l^-$, where $\mathbf{z}_l^- = (z_1^-, \cdots, z_d^-)^T$. Let $U = [u_{ij}]_{i,j=1,\cdots,d} = V^{+T} V^-$. Define the relative eigenvalues by $\tilde{\lambda}_{i,d}^+ = \lambda_{i,d}^+ / \Sigma_{i=1}^d \lambda_{i,d}^+$ and $\tilde{\lambda}_{j,d}^- = \lambda_{j,d}^- / \Sigma_{j=1}^d \lambda_{j,d}^-$. Then $(d\sigma_d \tau_d)^{-1} \Sigma_{i=1}^d (\dot{\mathbf{x}}_{ik}^+)(\dot{\mathbf{x}}_{il}^-)$ becomes

$$\begin{aligned} & (d\sigma_d \tau_d)^{-1} [z_1^+, \cdots, z_d^+] (\Lambda^+)^{\frac{1}{2}} V^{+T} V^- (\Lambda^-)^{\frac{1}{2}} [z_1^-, \cdots, z_d^-]^T \\ = \; & (\Sigma_{i=1}^d \lambda_{i,d}^+)^{-\frac{1}{2}} (\Sigma_{j=1}^d \lambda_{j,d}^-)^{-\frac{1}{2}} \Sigma_{s=1}^d \Sigma_{t=1}^d u_{s,t} z_s^+ z_t^- \sqrt{\lambda_{s,d}^+ \lambda_{t,d}^-} \\ = \; & \Sigma_{s=1}^d \Sigma_{t=1}^d u_{s,t} z_s^+ z_t^- \sqrt{\tilde{\lambda}_{s,d}^+ \tilde{\lambda}_{t,d}^-} \end{aligned}$$

The expectation of $\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_s^+z_t^-\sqrt{\tilde{\lambda}_{s,d}^+\tilde{\lambda}_{t,d}^-}$ is 0. Thus by Chebyshev's inequality,

$$\begin{aligned}
& Pr[|\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_s^+z_t^-\sqrt{\tilde{\lambda}_{s,d}^+\tilde{\lambda}_{t,d}^-}|\geq\varepsilon] \\
\leq\ & \varepsilon^{-2}E(\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_s^+z_t^-\sqrt{\tilde{\lambda}_{s,d}^+\tilde{\lambda}_{t,d}^-})^2 \\
=\ & \varepsilon^{-2}E[\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}\Sigma_{s'=1}^{d}\Sigma_{t'=1}^{d}u_{s,t}u_{s',t'}z_s^+z_{s'}^+z_t^-z_{t'}^-\sqrt{\tilde{\lambda}_{s,d}^+\tilde{\lambda}_{s',d}^+\tilde{\lambda}_{t,d}^-\tilde{\lambda}_{t',d}^-}] \\
=\ & \varepsilon^{-2}\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}^2\tilde{\lambda}_{s,d}^+\tilde{\lambda}_{t,d}^-
\end{aligned}$$

Since U is the product of two orthogonal matrix $U = V^{+T}V^-$, U is itself orthogonal. The relative eigenvalues satisfy the condition in Lemma 1. Thus by Lemma 1, $Pr[|\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_s^+z_t^-\sqrt{\tilde{\lambda}_{s,d}^+\tilde{\lambda}_{t,d}^-}|\geq\varepsilon]\rightarrow 0$, as $d\rightarrow\infty$. Thus $(d\sigma_d\tau_d)^{-1}\Sigma_{i=1}^{d}(\dot{\mathbf{x}}_{ik}^+)(\dot{\mathbf{x}}_{il}^-)$ converges to 0 in probability as $d\rightarrow\infty$. Further, since $\sigma_d^2\rightarrow\sigma^2<\infty$ and $\tau_d^2\rightarrow\tau^2<\infty$, $\frac{1}{d}\Sigma_{i=1}^{d}(\dot{\mathbf{x}}_{ik}^+)(\dot{\mathbf{x}}_{il}^-)\rightarrow 0$ in probability as $d\rightarrow\infty$.

The fourth term is the squared distance between the means, which is defined as $d\mu^2$.

The last term can be decomposed into two components: $\Sigma_{i=1}^{d}[E(\mathbf{x}_{i\cdot}^+)-E(\mathbf{x}_{i\cdot}^-)]\dot{\mathbf{x}}_{ik}^+$ and $\Sigma_{i=1}^{d}[E(\mathbf{x}_{i\cdot}^+)-E(\mathbf{x}_{i\cdot}^-)]\dot{\mathbf{x}}_{il}^-$. Let $\delta_i = E(\mathbf{x}_{i\cdot}^+)-E(\mathbf{x}_{i\cdot}^-)$. Note that $\Sigma_{i=1}^{d}\delta_i^2 = d\mu^2$. Each component, after being rescaled by d^{-1}, can be shown to converge to 0 in probability as $d\rightarrow\infty$. For example, the first component, rescaled by $(d\sigma_d)^{-1}$, becomes $\frac{1}{d\sigma_d}\Sigma_{i=1}^{d}[E(\mathbf{x}_{i\cdot}^+)-E(\mathbf{x}_{i\cdot}^-)]\dot{\mathbf{x}}_{ik}^+ = \frac{1}{d^{\frac{1}{2}}}\sqrt{\frac{1}{d\sigma_d^2}}\Sigma_{i=1}^{d}\delta_i\dot{\mathbf{x}}_{ik}^+ = \frac{1}{d^{\frac{1}{2}}}\Sigma_{i=1}^{d}\delta_i\Sigma_{s=1}^{d}v_{i,s}^+\sqrt{\tilde{\lambda}_{s,d}^+}z_s^+$. By Markov's inequality,

$$\begin{aligned}
& Pr(|\tfrac{1}{d^{\frac{1}{2}}}\Sigma_{i=1}^{d}\delta_i\Sigma_{s=1}^{d}v_{i,s}^+\sqrt{\tilde{\lambda}_{s,d}^+}z_s^+|>\varepsilon) \\
=\ & \varepsilon^{-2}E(\tfrac{1}{d^{\frac{1}{2}}}\Sigma_{i=1}^{d}\delta_i\Sigma_{s=1}^{d}v_{i,s}^+\sqrt{\tilde{\lambda}_{s,d}^+}z_s^+)^2 && = \varepsilon^{-2}\frac{1}{d}E(\Sigma_{s=1}^{d}\Sigma_{i=1}^{d}\delta_i v_{i,s}^+\sqrt{\tilde{\lambda}_{s,d}^+}z_s^+)^2 \\
=\ & \varepsilon^{-2}\tfrac{1}{d}\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_i v_{i,s}^+)^2\tilde{\lambda}_{s,d}^+E(z_s^+)^2 && = \varepsilon^{-2}\frac{1}{d}\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_i v_{i,s}^+)^2\tilde{\lambda}_{s,d}^+ \\
\leq\ & \varepsilon^{-2}\tfrac{1}{d}\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_i v_{i,s}^+)^2\max_i(\tilde{\lambda}_{i,d}^+) && = \varepsilon^{-2}\mu^2\max_i(\tilde{\lambda}_{i,d}^+)\rightarrow 0,\ \ as\ \ d\rightarrow\infty.
\end{aligned}$$

Note that $\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_i v_{i,s}^+)^2 = \Sigma_{i=1}^{d}\delta_i^2 = d\mu^2$ because V^+ is an orthogonal matrix, which keeps the norm of δ after transformation. Hence the first component $\Sigma_{i=1}^{d}[E(\mathbf{x}_{i\cdot}^+)-E(\mathbf{x}_{i\cdot}^-)]\dot{\mathbf{x}}_{ik}^+$, rescaled by d^{-1}, converges to 0 in probability as $d\rightarrow\infty$. And so does the second component $\Sigma_{i=1}^{d}[E(\mathbf{x}_{i\cdot}^+)-E(\mathbf{x}_{i\cdot}^-)]\dot{\mathbf{x}}_{il}^-$.

To summarize the analysis above, $\frac{1}{d}||\mathbf{x}_k^+-\mathbf{x}_l^-||^2\rightarrow\sigma^2+\tau^2+\mu^2,$ in probability, $as\ d\rightarrow\infty$.

Proof of Theorem 4:

Recall that the DWD hyperplane cut-off point P^* satisfies (13): $\frac{\alpha^*}{\beta^*} = \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}$. Let X_0^+ be a new data point from the $\mathcal{X}^+$-population. It was shown in Hall et al. (2005) that the rescaled squared distance of X_0^+ from O^+ and O^- are $\sigma^2(1+n_+^{-1})$ and $\mu^2+\sigma^2+\tau^2/n^-$ respectively, and it was known that the squared distance between O^+ and O^- was $\mu^2+\sigma^2/n^+ + \tau^2/n^-$. Let P be the projection of X_0^+ to the line O^+O^-, with distances to the two centroids being α and β, as diagrammed in Figure 8. It was shown by a series of geometric calculations in Hall et al. (2005) that $\frac{\alpha}{\beta} = \frac{\sigma^2/n^+}{\mu^2+\tau^2/n^-}$ when P lies on the real cut-off point P^*.

The point X_0^+ will be correctly classified as $\mathcal{X}^+$ type if it lies on the same side of the DWD hyperplane as O^+, i.e. if

$$\frac{\sigma^2/n^+}{\mu^2+\tau^2/n^-} < \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}. \tag{18}$$

It will be wrongly classified as $\mathcal{X}^-$ if

$$\frac{\sigma^2/n^+}{\mu^2+\tau^2/n^-} > \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}. \tag{19}$$

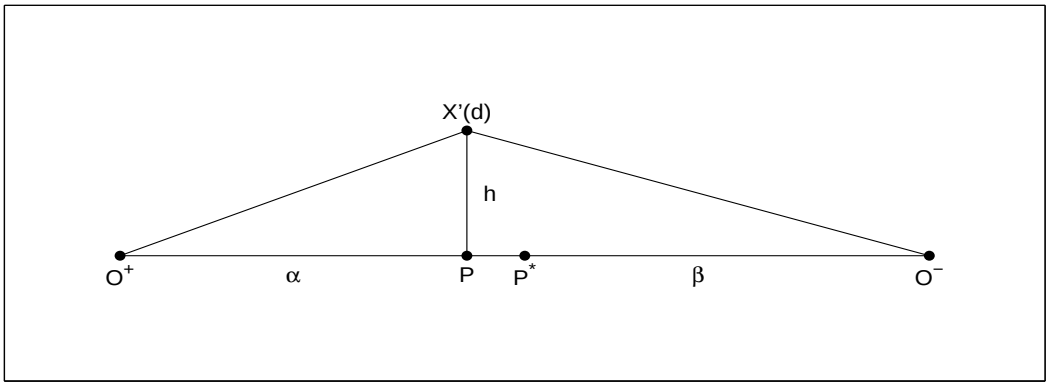

Figure 8: Projection point P of new data point $X'(d)$, DWD cut-off point P^*. α and β are distances of O^+ and O^- to P respectively.

The first and second parts of Theorem 4 follows (18) and (19) instantaneously. Now assume that $\sigma^2/[n_+^{\frac{3}{2}} w_+^{\frac{1}{2}}] \geq \tau^2/[n_-^{\frac{3}{2}} w_-^{\frac{1}{2}}]$. This ensures the non-negativity of $(n^- w_-/n^+ w_+)^{\frac{1}{2}}\sigma^2/n^+ - \tau^2/n^-$, the right hand side of the inequality in the first and second parts. Furthermore,

suppose that we have a data point X_0^- from the $\mathcal{X}^-$-population. By the inequality above, $\frac{\tau^2/n^-}{\sigma^2/n^+} \leq \left(\frac{w_- n^-}{w_+ n^+}\right)^{1/2}$.

Then for any positive μ^2 we have $\frac{\tau^2/n^-}{\mu^2+\sigma^2/n^+} < \frac{\tau^2/n^-}{\sigma^2/n^+} \leq \left(\frac{w_- n^-}{w_+ n^+}\right)^{1/2}$, i.e. X_0^- will always be classified as belonging to $\mathcal{X}^-$. Theorem 4 simply combines the analysis above.

Proof of Theorem 5:

Denote the centroids of the n^+-simplex from X^+ as $O_+^{n^+}$ and the n^--simplex from X^- as $O_-^{n^-}$. Also denote the population means of X^+ and X^- as O_+^∞ and O_-^∞ respectively. In the large d-limit, the expected squared distance, rescaled by d^{-1}, between $O_+^{n^+}$ and $O_-^{n^-}$ is $\mu^2 + \sigma^2/n^+ + \tau^2/n^-$. If we consider k more data vectors from X^+, the expected squared distance, rescaled by d^{-1}, between the centroids $O_+^{(n^++k)}$, of the new $(n^+ + k)-$simplex, and the centroid $O_-^{n^-}$, of the n^-−simplex is $\mu^2 + \sigma^2/(n^+ + k) + \tau^2/n^-$. Also the expected squared distance, rescaled by d^{-1}, between $O_+^{n^+}$ and $O_+^{(n^++k)}$ is $\left(\frac{k}{n^+(n^++k)}\right)\sigma^2$. This can be shown by calculating the distance between the two $(n^+ + k)-$dimensional vectors, $\sqrt{d}\sigma(\underbrace{n_+^{-1}, n_+^{-1}, \cdots, n_+^{-1}}_{n^+}, \underbrace{0, 0, \cdots, 0}_{k})^T$ and $\sqrt{d}\sigma(\underbrace{(n^+ + k)^{-1}, (n^+ + k)^{-1}, \cdots, (n^+ + k)^{-1}}_{n^+ + k})^T$, which are the centroids of the n^+-simplex $\{\sqrt{d}(\underbrace{1, 0, \cdots, 0}_{n^+}, \underbrace{0, \cdots, 0}_{k}), \cdots, \sqrt{d}(\underbrace{0, \cdots, 0, 1}_{n^+}, \underbrace{0, \cdots, 0}_{k})\}$ and the $(n^+ + k)$-simplex $\{\sqrt{d}(\underbrace{1, 0, \cdots, 0}_{n^+ + k}), \cdots, \sqrt{d}(\underbrace{0, \cdots, 0, 1}_{n^+ + k})\}$ respectively.

Thus by the Pythagorean theorem, $O_+^{n^+}O_+^{(n^++k)}$, $O_+^{n^+}O_-^{n^-}$ and $O_+^{(n^++k)}O_-^{n^-}$ form a right triangle, with $O_+^{n^+}O_-^{n^-}$ being the hypotenuse. And it follows that the angle between $O_+^{(n^++k)}O_-^{n^-}$ and $O_+^{n^+}O_-^{n^-}$ becomes approximately $cos^{-1}\left(\frac{\mu^2+\sigma^2/(n^++k)+\tau^2/n^-}{\mu^2+\sigma^2/n^++\tau^2/n^-}\right)^{\frac{1}{2}}$. Let $k \to \infty$. $O_+^{(n^++k)}$ converges to O_+^∞. Thus the angle between $O_+^\infty O_-^{n^-}$ and $O_+^{n^+}O_-^{n^-}$ becomes $cos^{-1}\left(\frac{\mu^2+\tau^2/n^-}{\mu^2+\sigma^2/n^++\tau^2/n^-}\right)^{\frac{1}{2}}$.

In the same manner, consider l more data vectors from X^-, and let $l \to \infty$. Then the angle between $O_+^\infty O_-^\infty$ and $O_+^{n^+}O_-^{n^-}$ is $cos^{-1}\left(\frac{\mu^2}{\mu^2+\sigma^2/n^++\tau^2/n^-}\right)^{\frac{1}{2}}$, i.e. the angle between the direction joining the means of two populations and the DWD direction joining the centroids of the n^+-simplex $\mathcal{X}^+(d)$ and the n^--simplex $\mathcal{X}^-(d)$ becomes $\theta = cos^{-1}\left(\frac{\mu^2}{\mu^2+\sigma^2/n^++\tau^2/n^-}\right)^{\frac{1}{2}}$.

References

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

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004), "Adjustment of Systematic Microarray Data Biases" *Bioinformatics* Vol. 20 No. 1, pages 105-114.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., Meyerson, M. (2001), "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses", *Proc Natl Acad Sci U S A* , 98(24):13790-5.

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

Duda, R., Hart, P. and Stork, D. (2001) *Pattern classification* (2nd edition), Wiley, New York.

Ge, N. and Simpson, D. G. (1998), "Correlation and High-Dimensional Consistency in Pattern Recognition", *Journal of the American Statistical Association* Vol. 93, 443, pp. 995-1006.

Hall, P., Marron, J. S. and Neeman, A. (2005), "Geometric Representation of High Dimension, Low Sample Size Data", *Journal of the Royal Statistical Society* Series B, 67, Part 3, pp. 427 - 444.

Jung, S. K. and Marron, J. S. (2008), "High Dimension, Low Sample Size PCA Consistency" *manuscript*.

Lin, Y., Lee, Y., and Wahba, G. (2002), "Support Vector Machine for Classification in Nonstandard Situation", *Machine Learning* 46, 191C202.

Liu, Y., Hayes, D. N., Nobel, A. and Marron, J. S. (2008), "Statistical Significance of Clus-

tering for High Dimension Low Sample Size Data", *Journal of the American Statistical Association*, forthcoming.

Marron, J. S., Todd, M. and Ahn, J. (2007), "Distance-Weighted Discrimination", *Journal of the American Statistical Association*, Vol.102, 480, pp. 1267-1271(5).

Meyerson, M. and Hayes, D. N. (2005), "Microarray Approaches to Gene Expression Analysis", *Molecular Diagnostics: For the Clinical Laboratorian. 2nd ed.* Tsongalis GJ, Coleman WB, eds. Totowa, NJ: Humana Press; 121-48.

Qiao, X. and Liu, Y. (2008), "Adaptive Weighted Learning for Unbalanced Multicategory Classification", *Biometrics*, forthcoming.

Schölkopf, B. and A.J. Smola (2002), *Learning with Kernels*, MIT Press, Cambridge, MA.

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