Weighted Distance Weighted Discrimination and its Asymptotic Properties

Xingye Qiao¹, Hao Helen Zhang², Yufeng Liu¹³, Michael J. Todd⁴, J. S. Marron¹

September 8, 2009

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. Several alternative classification criteria and the corresponding weighting choices are discussed. The second major contribution is substantial asymptotic study of weighted DWD. Let n be the sample size and d be the dimension of the data. Both high dimension low sample size asymptotics (d-asymptotics) and Fisher consistency of DWD are studied. The performance of weighted DWD is evaluated by simulated examples and two real data examples. The theoretical results are also confirmed by simulations.

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

 $^{^{1}\}mathrm{Department}$ of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599

 $^{^2\}mathrm{Department}$ of Statistics, North Carolina State University, Raleigh, NC 27695

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

⁴School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853

1 Introduction

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. Distance Weighted Discrimination (DWD) is a more recent classification method specifically designed for high dimension, low sample size (HDLSS) data settings, where the dimension d is greater than the sample size n.

The main idea behind SVM in the separable case is to find the separating hyperplane maximizing the total distance between the hyperplane and the closest data point of each class. Although SVM gives good performance in many real applications, it may suffer from a loss of generalization ability in HDLSS settings, as noted in Marron et al. (2007), due to the *data-piling* property, that is, the support vectors tend to pile up on top of each other at the boundaries of the margin when projected on the normal vector of the separating hyperplane. To overcome this data-piling issue, Marron et al. (2007) proposed DWD, which finds the hyperplane by minimizing the sum of the reciprocals of r_i , the distance from each data point to the hyperplane (min $\sum_i r_i^{-1}$). The improvement comes from the fact that DWD allows all the data vectors to have a direct impact on selecting the separating hyperplane, instead of only the support vectors as done by the SVM.

Although standard DWD (stdDWD) effectively avoids the data-piling problem in HDLSS settings, it was originally designed for balanced data, i.e., the case where the sample proportions for the two classes are similar. It has inefficient generalization ability under nonstandard situations, e.g., unequal costs or biased sampling (Lin, Lee and Wahba 2002), or when the two populations are seriously unbalanced (Qiao and Liu 2009). In particular, uneven class proportions can lead to a poor classifier which ignores the minority class. In this paper, we propose Weighted Distance Weighted Discrimination (wDWD) to incorporate class proportions as well as prior costs to improve upon standard DWD. In particular, wDWD uses the new objective function, $\min \sum_i w_i r_i^{-1}$, where w_i is the weight for the i^{th} training data point. Note that weighted DWD is more flexible than standard DWD by allowing flexible choices of weights. With appropriate choice of the weighting scheme, better generalization ability of weighted DWD follows under nonstandard situations. Optimal weighting schemes under different situations will be developed in Section 2.

Figure 1 shows a two-dimensional simulated example and the classification boundaries of the Bayes optimal classifier, wDWD and stdDWD. This example has unbalanced class proportions and unequal misclassification costs for the two classes. The boundary for wDWD is much closer to the Bayes optimal boundary than that of std-DWD. More details on this simulated example are provided in Section 3.2.



Figure 1: Two-dimensional unbalanced classification example: Comparison of stdDWD (dotted line) and wDWD (dashed line). Here the wDWD boundary is much closer to the Bayes optimal decision boundary (solid line).

Figure 2 shows the classification results for a high-dimensional simulated example (d = 1000; see more details in the constant signal case in Section 3.1.1). In the projection plot of all the data points on the stdDWD direction (top row of Figure 2), the stdDWD boundary (the vertical dashed line) works well for the training set (shown

as triangles). However, a potential problem is that it is too close to the positive class (on the right) because of the unbalanced class proportions. The test data (the + and \times signs) in Figure 2 show that stdDWD does not have good generalization ability. In the bottom row of Figure 2, note that the wDWD boundary provides a dramatic improvement over stdDWD for the test set.



Figure 2: High-dimensional simulated example: Projection plots of data points on the stdDWD (top) and wDWD (bottom) directions. The separating hyperplanes intersect the wDWD and stdDWD directions at the two dashed vertical lines respectively. These plots show much better performance of wDWD in this HDLSS setting.

We develop asymptotic properties of the wDWD classifier in HDLSS settings in Section 4.1. Ge and Simpson (1998) analyzed the high-dimensional asymptotics of some classifiers. HDLSS data have a special structure which gives insight into the classification problem. Hall et al. (2005) showed that under certain conditions, there exists a *geometric representation* of HDLSS data, which implies that the pairwise distances between the n^+ (n^- resp.) data points from the same class "+1" ("-1" resp.) are approximately constant as $d \to \infty$ with n^+ (n^- resp.) fixed. As a consequence, each sample from one class (of size n^+ or n^-) can be viewed as an regular ($n^+ - 1$) (($n^- - 1$) resp.)-simplex. The results in Hall et al. (2005) assume the entries of each data vector to be *nearly independent*, in the sense that when they are viewed as a time series with the time index d, these entries must satisfy a ρ -mixing condition. Ahn et al. (2007) extended their work by showing that the conditions can be relaxed to asymptotic properties of the sample covariance matrix and its eigenvalues. However, a counterexample due to John Kent (discussed in Section 4.1.1) suggests that an additional Gaussian assumption is needed. A much broader set of assumptions for geometric representation has been developed in Jung and Marron (2008). In this article, our theory makes use of this broader framework.

To study asymptotic properties of wDWD, we develop a geometric representation for two data samples from two classes as in Hall et al. (2005) but under milder assumptions. Using the framework built in Section 4.1.2, we study the asymptotic properties of two aspects of wDWD as $d \to \infty$ with the sample size n fixed. Both properties follow from the geometric representation described above. First, we study the classification error of wDWD. Second, we explore the relationship between the wDWD direction and the optimal linear classification direction. Both aspects are driven by appropriate notions of signal to noise ratios, defined in terms of class means and within-class variances. These properties are verified by simulations in Section 4.2.

As another theoretical contribution, we prove Fisher consistency for wDWD in Section 4.3 through an equivalent formulation of the optimization problem for wDWD.

The rest of this article is organized as follows. We propose weighted DWD in Section 2, focusing on optimal weighting schemes under different scenarios. Numerical studies are given in Section 3 based on simulated and real data examples. In Section 4.1, we provide the geometric representation of two HDLSS data samples from two classes and study the HDLSS asymptotic properties of wDWD, followed by a simulation confirmation in Section 4.2. Fisher consistency of weighted DWD is provided in Section 4.3. Some concluding remarks are given in Section 5. Proofs of the theoretical results are included in the Appendix.

2 Weighted DWD

2.1 General Classification Problems

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

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

A linear classifier $\phi(\mathbf{x})$ can be obtained from $\phi(\mathbf{x}) = \operatorname{sign}(f(\mathbf{x}))$, where $f(\mathbf{x}_i) = f_i = \mathbf{x}'_i \boldsymbol{\omega} + b$. The data vector with covariate \mathbf{x}_i is classified to Class "+1" if $\operatorname{sign}(f_i) = +1$ and Class "-1" otherwise.

2.2 Formulation for Weighted DWD

Suppose the classification boundary is represented as a separating hyperplane, $\mathbf{x}'\boldsymbol{\omega} + b = 0$. The standard DWD proposed in Marron et al. (2007) seeks to find a separating hyperplane minimizing a notion of inverse distance between each point and the hyperplane (details below). As mentioned in Section 1, standard DWD has some limitations for unbalanced data. For example, in Figure 1, the stdDWD classification boundary is pushed away from the Bayes boundary, mainly caused by the dramatic difference between two class proportions. Our proposed weighted DWD aims to address this problem by allowing flexible weights for data points from different classes. In particular, wDWD solves ($\boldsymbol{\omega}, b$) via the following optimization problem,

$$\min_{\boldsymbol{\omega},b,\boldsymbol{\xi}} \sum_{i=1}^{n} W(y_i) \Big(\frac{1}{r_i} + C\xi_i \Big), \tag{2}$$

s.t. $r_i = y_i(\mathbf{x}'_i \boldsymbol{\omega} + b) + \xi_i$, $\boldsymbol{\omega}' \boldsymbol{\omega} \leq 1$, $r_i \geq 0$, $\xi_i \geq 0$ for $i = 1, 2, \dots, n$. (3) Here we assign different weights to data vectors from different classes. Note that the solution to (2) is totally determined by the ratio of W(+1) and W(-1), instead of the exact values of the two weights. The standard DWD is a special case of the weighted DWD with equal weights, W(+1) = W(-1).

To have a better understanding of (2), we first consider a simple separable setting with a choice of C where all ξ_i 's are 0. Then wDWD minimizes the total weighted inverse distances of all points to the decision boundary. When the perfect separation is not possible, (2) allows violation with amount ξ_i for training data point *i*.

The constant parameter C in (2) controls the penalty on the variable ξ_i , the amount of violation of classification. Note that C plays the similar role as the tuning parameter in the SVM (see eq. (54) in Chen et al. 2005; also see Vapnik 1995, Schölkopf and Smola 2002). This optimization problem in (2) can be reformulated as a second-order cone programming (SOCP) problem (Alizadeh and Goldfarb 2003), as shown in Marron et al. (2007).

Define $W(-1)I[y = -1]I[\phi(\mathbf{x}) = +1] + W(+1)I[y = +1]I[\phi(\mathbf{x}) = -1]$ as the weighted 0-1 loss function corresponding to problem (2). The Bayes optimal decision rule for this weighted 0-1 loss is

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

More details will be provided in Section 4.3.

2.3 Optimal Weighting Schemes

In this section, we discuss two nonstandard classification situations which are commonly encountered in practice, and study the choices of optimal weights for each situation. We consider the situation of unequal costs in Section 2.3.1 and the biased sampling situation in Section 2.3.2. An alternative classification criterion, *Mean Within Group Error* (MWGE), is also discussed.

2.3.1 Unequal Costs

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

		Classify as	
		+1	-1
True nonulation.	+1	0	c^{-}
The population:	-1	c^+	0

Table 1: Unequal costs for different types of misclassification.

Using the Overall Misclassification criterion (OM), 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 given $\mathbf{X} = \mathbf{x}$, is $E[L(\phi)|\mathbf{X} = \mathbf{x}] = c^+[1 - p(\mathbf{x})]I[\phi(\mathbf{x}) = +1] + c^- p(\mathbf{x})I[\phi(\mathbf{x}) = -1]$. The Bayes optimal decision rule ϕ^* for this loss function minimizes the risk and is given by

$$\phi^{*}(\mathbf{x}) = \begin{cases} +1 & \text{if } \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} > \frac{c^{+}}{c^{-}} \\ -1 & \text{if } \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} < \frac{c^{+}}{c^{-}}, \end{cases} \text{ or } \phi^{*}(\mathbf{x}) = \text{sign}[p(\mathbf{x}) - \frac{c^{+}}{c^{+} + c^{-}}]. \tag{5}$$

Comparing this to (4), by defining $W(+1) = c^{-}$ and $W(-1) = c^{+}$, we have the two Bayes rules identical with each other.

Our discussions so far assume the classic Overall Misclassification rate criterion (OM). This criterion has some limitations. For example, if the two classes are extremely unbalanced, a naive classifier, which classifies all the data vectors to the majority class, can still be regarded as a good one by this criterion. Alternatively, one can use the Mean Within Group Error (MWGE) criterion (Qiao and Liu 2009). This criterion considers the average of the within-class errors. Under MWGE, the modified 0-1 loss function becomes $\frac{c^+}{\pi^-}I[y=-1]I[\phi(\mathbf{x})=+1] + \frac{c^-}{\pi^+}I[y=+1]I[\phi(\mathbf{x})=-1]$. The corresponding Bayes rule ϕ_* is given by $\phi_*(\mathbf{x}) = sign[p(\mathbf{x}) - \frac{c^+/\pi^-}{c^+/\pi^-+c^-/\pi^+}]$, which implies that the choice of the weighting scheme under MWGE is $W(+1) = \frac{c^-}{\pi^+}, W(-1) = \frac{c^+}{\pi^-}$. Discussion on several other alternative criteria will be given in Section 2.4.

2.3.2 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. Because we build the classification model on the sample while we predict a future data vector from the population, the discrepancy of the class proportion ratios between the sample and the population could lead to a problematic classifier. Lin, Lee and Wahba (2002) discussed nonstandard situations for the SVM.

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 labeled as in Table 2. Let (\mathbf{X}_s, Y_s) be a random pair that has the same distribution as the sample. Note that the conditional densities g_s^+ and g_s^- are the same as g^+ and g^- . Then the conditional probability of a case from the sample belonging to the +1 class given that $\mathbf{X}_s = \mathbf{x}$ is

$$p_s(\mathbf{x}) = \Pr(Y_s = +1 | \mathbf{X}_s = \mathbf{x}) = \frac{\pi_s^+ g_s^+(\mathbf{x}_s)}{\pi_s^+ g_s^+(\mathbf{x}_s) + \pi_s^- g_s^-(\mathbf{x}_s)} = \frac{\pi_s^+ g^+(\mathbf{x})}{\pi_s^+ g^+(\mathbf{x}) + \pi_s^- g^-(\mathbf{x})}.$$
 (6)
Comparing (1) and (6), the relationship of the odds ratio of $p(\mathbf{x})$ from the population

Comparing (1) and (6), the relationship of the odds ratio of $p(\mathbf{x})$ from the population and that of $p_s(\mathbf{x})$ from the sample is

$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \frac{\pi^+ g^+(\mathbf{x})}{\pi^- g^-(\mathbf{x})} = \frac{\pi_s^+ g^+(\mathbf{x})}{\pi_s^- g^-(\mathbf{x})} \frac{\pi^+ \pi_s^-}{\pi^- \pi_s^+} = \frac{p_s(\mathbf{x})}{1-p_s(\mathbf{x})} \frac{\pi^+ \pi_s^-}{\pi^- \pi_s^+}$$

Then the Bayes rule in (5) can be expressed in terms of $p_s(\mathbf{x})$ as

$$\phi^*(\mathbf{x}) = \begin{cases} +1 & \text{if } \frac{p_s(\mathbf{x})}{1 - p_s(\mathbf{x})} > \frac{c^+ \pi_s^+ \pi^-}{c^- \pi_s^- \pi^+} \\ -1 & \text{if } \frac{p_s(\mathbf{x})}{1 - p_s(\mathbf{x})} < \frac{c^+ \pi_s^+ \pi^-}{c^- \pi_s^- \pi^+}, \end{cases} \text{ or } \phi^*(\mathbf{x}) = \text{sign}[p_s(\mathbf{x}) - \frac{c^+ \pi^- / \pi_s^-}{c^+ \pi^- / \pi_s^- + c^- \pi^+ / \pi_s^+}]. \end{cases}$$

Note that because the calculation of a classifier is based on the sample, instead of the population, when biased sampling exists, $p_s(\mathbf{x})$ should be used in the classification rule $\phi(\mathbf{x})$ whereas $p(\mathbf{x})$ in (5) is not useful, since $p(\mathbf{x}) \neq p_s(\mathbf{x})$. Again, using the formulation in (4), we can see that the choice of weights becomes $W(+1) = \frac{c^-\pi^+}{\pi_s^+}$ and $W(-1) = \frac{c^+\pi^-}{\pi_s^-}$.

Now we consider the situation where the MWGE criterion is used. The Bayes rule ϕ_* under MWGE is then given by

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

Accordingly, we can define the weights $W(+1) = \frac{c^-}{\pi_s^+}, W(-1) = \frac{c^+}{\pi_s^-}.$

In summary, the optimal weighting scheme is displayed in Table 3.

Criterion	OM	MWGE
W(+1)	$\frac{c^-\pi^+}{\pi_s^+}$	$\frac{c^-}{\pi_s^+}$
W(-1)	$\frac{c^+\pi^-}{\pi_s^-}$	$\frac{c^+}{\pi_s^-}$

Table 3: Optimal weighting schemes under two criteria.

2.3.3 Parameter Tuning for wDWD

Marron et al. (2007) discussed the choice of C and suggested that C should be a large constant (for example 100 in their work) divided by a notion of *typical squared distance* of the training points (for example squared median of the pairwise interclass Euclidean distances). The usage of typical squared distance will result in a choice of Cthat is essentially "scale-invariant". From the simulation results in Section 3.1, where we tune for the best parameter C using a grid search on the tuning set, we find that the tuned C values are reasonable close to their suggestion.

It is worth noting that careful tuning needs to be done for DWD when the data

are unbalanced and the signal (denoted by the distance between the two population means) is small. In particular, a small C should be avoided. For unbalanced data, a small value of C tends to yield undesired results for stdDWD, with most data vectors classified into the majority class. This is because DWD optimization avoids large values of reciprocal distances $1/r_i$ by sacrificing the data from the minority class. Thus Cneeds to be large enough to increase the misclassification cost. Weighted DWD, on the other hand, alleviates this problem in tuning since the adverse effect of the unbalanced proportion ratio on standard DWD can be greatly reduced if the weighting scheme is appropriately chosen.

2.4 Alternative Criteria and Adaptive Weighting

In Section 2.3, we introduced the optimal weighting schemes under the OM and MWGE criteria (Table 3). Recall that the OM criterion aims to minimize the OM cost. Qiao and Liu (2009) pointed out that this criterion may result in a high error for the minority class when the proportions are unbalanced. In addition to MWGE, they introduced Mean Square Within Group Error (MSWGE). In this paper, we also consider the criterion of Maximal Within Group Error (MaxWGE). Let $e_j = E[I(\phi(\mathbf{X}) \neq j)|Y = j]$ be the conditional error for class j. We reformulate the minimization of these criteria equivalently as follows:

(i) *OM*:

$$\arg\min_{\phi}\pi^+e_+ + \pi^-e_-;$$

(ii) the alternatives:

$$\arg\min_{\phi} (e_{+}^{p} + e_{-}^{p})^{\frac{1}{p}} = \begin{cases} \arg\min_{\phi} \frac{1}{2}(e_{+} + e_{-}) \ (MWGE), \ if \ p = 1, \\ \arg\min_{\phi} \sqrt{\frac{1}{2}(e_{+}^{2} + e_{-}^{2})} \ (MSWGE), \ if \ p = 2, \\ \arg\min_{\phi} \max(e_{+}, e_{-}) \ (MaxWGE), \ if \ p = \infty. \end{cases}$$
(7)

The alternative criteria focus on $|\mathbf{e}|_p$, the L_p norm of the within-class error vector $\mathbf{e} = [e_+, e_-]^T$. An important feature of MWGE, MSWGE and MaxWGE is that they do not require knowledge of, or even specification of, the prior proportions π^+ and π^- .

Thus, these criteria overcome the severe limitations of OM in the unbalanced case. The three alternative criteria provide different summaries of the error. The MWGE (L_1) criterion tends to minimize the mean of the within-class errors while the MSWGE (L_2) criterion minimize the mean and variation at the same time. The MaxWGE (L_{∞}) criterion controls the worse class error. Choice among these will depend on the statistical context at hand.

To demonstrate these different criteria, we consider a one-dimensional toy example with two classes, the density curves of which are two triangles as shown in Figure 3. Note that the OM Bayes rule is sensitive to the change of the class proportions and is not desirable when the class proportions are unknown. On the other hand, the Bayes rules for the alternative criteria do not change with proportions. Different alternative criteria provide different Bayes cut-off points in this example.



Figure 3: One-dimensional density curves for two populations and the Bayes rules for OM (dotted), MWGE (dashed), MSWGE (solid) and MaxWGE (dot-dashed) criteria when the population proportion ratio is 5:1, 3:1, 1:1, 1:3 or 1:5. Shows OM is very sensitive to class proportions, and compares the three alternative criteria.

Qiao and Liu (2009) showed that there exist closed forms for the OM and MWGE Bayes rules, which lead to the optimal DWD weighting schemes introduced in Section 2.3. However, the Bayes rules under the other alternative criteria do not seem to have simple closed forms. Therefore, in order to achieve better results based on the alternative criteria, we propose a two-step procedure to adaptively choose the weights using the sample within-class errors. The proposed adaptive procedure is implemented as follows:

(i) Train wDWD with the MWEG optimal weights (Table 3), $W(\pm 1)$. Calculate the within-class errors \hat{e}_+ and \hat{e}_- for the combined dataset including both training and tuning sets.

(ii) Update weights for class j as $W(j) \cdot exp(\hat{e}_j)$, for $j \in \{+, -\}$, and calculate wDWD using the new weights. We call the resulting classifier the adaptive weighted DWD (awDWD).

The adaptive weighting adjustment at the second step gives a bigger weight to the class with larger error. We will show in Section 3.1.2 that awDWD can provide additional improvement over wDWD.

3 Numerical Study

In this section, we compare wDWD with stdDWD and several other classification methods, based on two simulation studies (two high-dimensional examples (independent predictors and correlated predictors) in Section 3.1 and a low-dimensional example in Section 3.2) and two real data examples (Section 3.3).

We consider L_1 SVM (Fung et al. 2004), weighted SVM (wSVM), standard SVM (stdSVM), the L_1 penalized logistic regression (L_1 Logit; Lokhorst 1999, Shevade and Keerthi 2003) and the L_2 penalized logistic regression (L_2 Logit; Lee and Silvapulle 1988, Le Cessie and Van Houwelingen 1992). L_1 SVM and L_1 Logit use the L_1 penalty for variable selection. Weighted SVM is the weighted version of standard SVM, where we use the same weighting scheme as that of weighted DWD. In Section 3.1.1, we also implement awDWD to show its performance. For comparison purpose, we apply

the same adaptive weights for wSVM, namely awSVM. As a remark, we note that the results for the L_1 and L_2 Logit are not available for some examples due to numerical difficulties.

3.1 Simulation: High-dimensional Example

Let the dimension d = 1000, and the sample size of the training data n = 200. Assume that the data are balanced with $\pi^+ = \pi^- = 50\%$ and equal costs $c^- = c^+$, but with a biased sampling, $\pi_s^+ = 20\%$ and $\pi_s^- = 80\%$. For simplicity, we denote w_+ and w_- as the two weights W(+1) and W(-1). The weights for this data set are $w_+ = 2.5$, and $w_- = 0.625$. Note that because $\pi^+ = \pi^-$, the two weighting schemes given by Table 3 and the two Bayes rules for the two criteria (OM and MWGE) are the same.

3.1.1 Independent Predictors

We consider three settings of high-dimensional simulated data, namely constant signal, proportional signal and sparse signal. In the constant signal setting, the variablewise mean differences are equal for all 1000 variables, while in the sparse signal setting, only the first 10 variables have nonzero mean differences. One intermediate setting is the proportional signal where the squared mean difference for each variable is proportional to the variable index ($\{1, ..., 1000\}$). The data vectors from the positive class follow d-dimensional normal distributions $N_d(u_1\mathbf{1}_d, 0.75^2\mathbf{I}_d)$, $N_d(u_2(1, 2, ..., d)^T, 0.75^2\mathbf{I}_d)$ and $N_d(u_3(\mathbf{1}_{10}^T, 0, ..., 0)^T, 0.75^2\mathbf{I}_d)$ corresponding to the three settings, where $\mathbf{1}_k =$ $[1, 1, ..., 1]^T$ is the k-dimensional vector of 1's. The negative data vectors are generated in a similar manner except with negative means $-u_1\mathbf{1}_d$, $-u_2(1, 2, ..., d)^T$ and $-u_3(\mathbf{1}_{10}^T, 0, ..., 0)^T$ in the normal distributions. The positive constants u_1, u_2 and u_3 are chosen so that the Euclidean distances of the two population means for the three settings are all equal to 3. For tuning and testing purposes, we generate a tuning set with size 200 and a test set with size 600. We replicate this simulation 100 times.

From Table 4, we first compare the non-adaptive methods for the three settings.

Type		Constant		Proportio		ortional		Sparse	
Criteria	OM/MWGE	MSWGE	MaxWGE	OM/MWGE	MSWGE	MaxWGE	OM/MWGE	MSWGE	MaxWGE
Bayes	2.07(0.18)	2.12(0.17)	2.43(0.15)	2.38(0.25)	2.45(0.25)	2.83(0.3)	1.93(0.24)	1.96(0.24)	2.2(0.25)
wDWD	13.52(0.89)	14.96(1.46)	19.23(2.36)	17.62(3.76)	21.7(5.78)	29(8.49)	16.42(1.04)	20.62(1.77)	28.57(2.72)
awDWD	12.57(0.57)	13.04(0.71)	15.2(1.31)	13.7(0.93)	15.04(1.52)	19.23(2.43)	13.38(0.4)	14(0.49)	16.73(0.97)
stdDWD	45.8(0.46)	64.72(0.69)	91.53(0.97)	46.05(0.34)	65.1(0.48)	92.07(0.68)	45.1(0.42)	63.78(0.59)	90.2(0.84)
L_1 SVM	35.07(0.9)	35.4(0.9)	39.37(1.09)	32.72(0.93)	38.57(1.77)	52.37(3.03)	7.38(0.31)	9.08(0.36)	12.63(0.5)
wSVM	22.62(0.95)	30.93(1.5)	43.7(2.13)	24.55(1.59)	33.41(2.45)	47.17(3.5)	19.63(1.38)	25.66(2.14)	36(3.12)
awSVM	16.82(0.95)	21.44(1.54)	29.93(2.3)	17.25(0.91)	22.16(1.53)	31.03(2.25)	15.15(0.74)	17.8(1.23)	24.2(1.94)
stdSVM	29.3(0.51)	41.11(0.74)	58.13(1.04)	30.22(0.73)	42.15(1.07)	59.6(1.51)	30.42(0.77)	42.25(1.16)	59.73(1.64)

Table 4: Summary statistics of the simulation results for the three simulation settings: Averaged OM/MWGE, MSWGE, and MaxWGE (in percentage) over 100 runs. The numbers reported in the parentheses are the standard error.

In each setting, wDWD works much better than stdDWD. In addition, wDWD works better than all the other non-adaptive methods in the constant signal and proportional signal settings. For the sparse signal case, both L_1 SVM and L_1 Logit are better than wDWD. This is expected since our current wDWD does not attempt to handle sparsity by variable selection. A potential approach to improving wDWD for the sparse signal setting is to design a classification algorithm combining wDWD and some sparse penalty such as the L_1 (Tibshirani 1996) or SCAD (Fan and Li 2001) penalty to implement variable selection.

Table 4 also indicates that adaptive weighted DWD introduced in Section 2.4 works very well. In all three signal settings, awDWD dominates all the other methods except L_1 SVM in the sparse setting. It seems that the advantage of awDWD comes from the fact that it prevents wDWD from overweighting by incorporating both class proportions and within-class performance in the weights. Moreover, both adaptive weighting methods (awDWD and awSVM) provide further improvement on wDWD and wSVM in these examples, in terms of the MSWGE and MaxWGE criteria, in addition to the MWGE criterion.

We note that for the non-adaptive weighting methods, even though their OM error or MWGE seem to be fine, their MSWGE and MaxWGE are not satisfactory (e.g., wDWD for proportional signal has MaxWGE of 29%). Adaptive weighting methods usually lead to lower MSWGE and MaxWGE as shown in Table 4.



Figure 4: Projection plots of all the data vectors to the two-dimensional space spanned by the Bayes optimal classification direction (Bayes drn) and the wDWD direction (in the left panels) or the L_1 SVM direction (in the right panels) for simulated data from the constant signal setting (the first row) and the sparse signal setting (the second row).

Among these different methods, L_1 SVM is the one that performs much better than wDWD under the sparse signal setting. To further compare them, we consider their classification directions. Figure 4 contains four projection plots which study the angles between the optimal linear classification direction and the classification direction from wDWD (in the left panel) or from L_1 SVM (in the right panel) for the constant signal setting (in the first row) or the sparse signal setting (in the second row). We can see that the angles for wDWD are comparable between the two settings, whereas the angles for L_1 SVM are larger than those for wDWD in the constant signal setting but smaller in the sparse signal setting. These angles help to explain the difference between classification performances of these two methods. Note that there is severe data-piling for L_1 SVM, as shown in the right column of Figure 4.

3.1.2 Correlated Predictors

We modify the high-dimensional example in Section 3.1.1 by adding correlations among the predictors. Instead of assuming i.i.d. Gaussian noise, we let the noise term be an autoregressive process of order 1 (AR(1)) with marginal variance 0.75^2 . We use several choices of the autocorrelation parameter, $\rho = 0.05$, 0.35, 0.65 and 0.95. Before adding the three types of variablewise mean difference (which was chosen for each case to give good separation between the classifiers, while conveying the challenge of highly correlated errors), we permute the order of the variables to break down the AR structure. We plot the OM test errors in Figure 5.



Figure 5: Simulation results of wDWD (solid), stdDWD (wide-dotted), wSVM (dashed), stdSVM (dotted) and L_1 SVM (dot-dashed) for three 1000-dimensional settings (constant, proportional and sparse signals) with AR(1) noise where $\rho = 0.05, 0.35, 0.65$ and 0.95.

For all three signal settings, wDWD works the best when $\rho = 0.05$ and 0.35 except for L_1 SVM in the sparse setting. For larger ρ , such as 0.65 and 0.95, wDWD and wSVM are comparable. In the sparse setting, L_1 SVM is the best as expected. An important point is that wDWD is less efficient in the highly correlated case, which was also noted by Ahn and Marron (2009).

In these studies, we choose the tuning parameter C based on a search grid of

 $10^{\{-4,-3.5,...3.5,4\}}$. In all the three settings, we observe that our tuning parameter search procedure tends to choose $10^{-1.5}$ for weighted DWD, while the recommendation of C by Marron et al. (2007) turns out to be about $10^{-1.05}$. Based on our limited experience, their recommendation appears to work reasonably well.

3.2 Simulation: Low-dimensional Example

In this section, we consider a two-dimensional simulated example. The underlying population has unbalanced class proportions with $\pi^+ = 10\%$ and $\pi^- = 90\%$. The sampling distribution is based on 400 observations with proportions $\pi_s^+ = 40\%$ and $\pi_s^- = 60\%$. For the purpose of tuning, we randomly divide each dataset into two halves: one for training and the other for tuning. Thus we have $n^+ = 80$, $n^- = 120$ for each set. As in Section 3.1, we choose the tuning parameter by a grid search.

Suppose that the data vectors from the positive class follow a bivariate normal distribution $N_2((0,0)^T, I_2)$, whereas the negative class is $N_2((1.5, 1.5)^T, I_2)$. The size of the test dataset is 600 ($n^+ = 60$, $n^- = 540$). Furthermore, we set the cost of a false negative classification to be twice of the cost of a false positive one, i.e., $c^- = 2c^+$. We replicate this simulation 100 times. Due to the extreme unbalanced class proportion ratio (1:9), we choose to use the MWGE criterion (Qiao and Liu 2009) and the weighting scheme under the MWGE given by Table 3. Here the Bayes rule ϕ_* is $sign[1.5 + \frac{1}{3} \log 2 - x_1 - x_2]$.

	Trainin	g Data	Test Data			
Classifier	False negative	False positive	False negative	False positive	MWG cost	
(%)	$c^{-} = 2$	$c^{+} = 1$	$c^- = 2$	$c^{+} = 1$		
Bayes	11.29(0.34)	17.87(0.378)	11.73(0.438)	18.71(0.186)	21.00(0.449)	
wDWD	7.2(0.218)	24.95(0.558)	7.92(0.417)	26.49(0.485)	21.16(0.388)	
stdDWD	18.95(0.401)	9.78(0.242)	20.87(0.580)	10.61(0.203)	26.17(0.543)	
wSVM	6.99(0.239)	25.14(0.530)	7.68(0.403)	26.69(0.428)	21.03(0.387)	
stdSVM	18.65(0.445)	10.11(0.249)	20.65(0.562)	10.81(0.239)	26.05(0.523)	

Table 5: Training and test error for the low-dimensional simulation: mean misclassification rates for each class and the Mean Within Group cost (MWG cost), on both training and test data, over 100 replications. The numbers in the parentheses are standard error.

The results of the simulation are reported in Table 5. Both weighted algorithms (wDWD and wSVM) are better than their standard counterparts. However, in contrast to the better performance of wDWD in high-dimensional settings, here wDWD and wSVM perform similarly. We have shown an illustrating plot in Figure 1 (Section 1), where the wDWD boundary is much closer to the Bayes boundary than that of the stdDWD.

3.3 Real Data Examples

In this section we demonstrate the improved performance from weighting DWD on two data examples, the Human Lung Carcinomas Micro-arrays Dataset (Lung cancer data) (Bhattacharjee et al. 2001; http://www.broad.mit.edu/mpr/lung/) and the Gisette data (http://www.nipsfsc.ecs.soton.ac.uk/).

The Lung cancer data set has six classes: adenocarcinoma, squamous, pulmonary carcinoid, colon, normal and small cell carcinoma, with sample sizes of 128, 21, 20, 13, 17 and 6 respectively. Liu et al. (2008) used this data as a test set to demonstrate their proposed significance analysis of clustering. We combine the last four and the first two subclasses to form the positive and negative classes respectively. We randomly split the data into training $(n_{+} = 100 \text{ and } n_{-} = 40)$ and test (49+16) sets.

The context of the Gisette dataset is a handwritten digit recognition problem to separate the highly confusable digits '4' and '9'. The original dataset has 6000 (3000+3000)cases in the training set and 1000 (500+500) in a separate test set. We randomly choose 600 and 200 cases for each class from the original training set, and equally split them to form the new training and tuning set. There are 5000 predictors in all, where 2500 predictors have true predictive power and the rest of them are deliberately irrelevant.

For the choice of the tuning parameter C, we use 5-fold cross validation for the Lung cancer data and use the tuning set for the Gisette data. For computational simplicity, we use the MWGE weighting scheme in Table 3.

	Lung cancer data				Gisette data		
	OM	MWGE	MSWGE	MaxWGE	OM/MWGE	MSWGE	MaxWGE
wDWD	5.11(0.25)	4.88(0.28)	5.66(0.29)	7.26(0.38)	11.48(0.45)	12.86(0.53)	17(0.88)
stdDWD	7.49(0.26)	10.93(0.43)	13.21(0.57)	18.03(0.83)	15.54(0.5)	20.16(0.76)	28.36(1.09)
L_1 SVM	7.88(0.25)	9.43(0.39)	10.68(0.46)	13.87(0.66)	12.74(0.66)	14.37(0.74)	19.3(1.04)
wSVM	4.91(0.25)	5(0.29)	5.75(0.3)	7.4(0.39)	11.49(0.44)	13.03(0.66)	17.32(1.15)
stdSVM	6.03(0.25)	7.64(0.41)	8.94(0.51)	11.78(0.72)	14.12(0.69)	17.22(1.02)	23.92(1.49)

Table 6: Summary statistics of the classification errors in the Lung cancer data and the Gisette data: Mean classification errors (OM, MWGE, MSWGE, and MaxWGE) for the test sets over 100 random splitting of training and test sets. The numbers reported in the parentheses are the standard error.

We run the random splitting 100 times and report the mean of the errors for the test data, and the associated standard error, in Table 6. For both data, weighted DWD appears to be better than stdDWD, L_1 SVM, stdSVM, and L_1 Logit for all types of criteria. For the Lung cancer data, the weighted DWD works better than wSVM for the MWGE, MSWGE and MaxWGE, although not for the OM error. For the Gisette data, the weighted DWD works slightly better than weighted SVM for all criteria.

4 Theoretical Results

In this section, we study several theoretical aspects of wDWD. HDLSS asymptotics are discussed in Section 4.1, followed by simulation validation in Section 4.2. Fisher consistency for wDWD is discussed in Section 4.3.

4.1 HDLSS Asymptotics for Weighted DWD

In this section, we explore the HDLSS asymptotics of wDWD. We will first improve the theory of Ahn et al. (2007) using a much broader set of assumptions. In addition, we geometrically represent *two* data samples under the new assumption.

4.1.1 Geometric Representations for One Sample under Mild Conditions

First consider the positive class $\mathcal{X}^+(d) = \{\mathbf{x}_1^+(d), \mathbf{x}_2^+(d), \dots, \mathbf{x}_{n^+}^+(d)\}$ 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 def-

inite 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 diagonal matrix of eigenvalues. Furthermore, 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 $S_{D,d}^+ = d^{-1} \mathbf{X}_d^+ \mathbf{X}_d^+$, reversing the roles of rows and columns in the data matrix. 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,

$$dS_{D,d}^{+} = \sum_{i=1}^{a} \lambda_{i,d}^{+} W_{i,d}^{+}.$$
(8)

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

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,

(i) Each column of \mathbf{X}_d^+ has zero mean and positive definite covariance matrix Σ_d^+ .

(ii) The fourth moment of each entry of each column is uniformly bounded by $M^+ > 0$ and also the representation in (8) holds for each \mathbf{X}_d^+ .

(iii) 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. (iv) The eigenvalues of Σ_d^+ are sufficiently diffused, in the sense that

the eigenvalues of
$$\Sigma_d$$
 are sufficiently diffused, in the sense that $\epsilon^+ = \frac{\Sigma_{i=1}^d (\lambda_{i,d}^+)^2}{1 + 1 + 1 + 2} \longrightarrow 0$ as $d \to \infty$

$$\epsilon_d^+ = \frac{\sum_{i=1}^{d} (\lambda_{i,d})}{(\sum_{i=1}^{d} \lambda_{i,d}^+)^2} \to 0 \quad \text{as} \quad d \to \infty.$$
(9)

(v) The sum of the eigenvalues of Σ_d^+ is the same order as d, in the sense that $\sigma_d^2 = O(1)$ and $1/\sigma_d^2 = O(1)$.

Condition (9) can be viewed as a measure of the sphericity of the data matrix. This 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 ρ -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 (in particular, condition (iii) did not appear there), the sample eigenvalues behave as if they follow an identity covariance matrix, in the sense that $\frac{1}{\sigma^2}S_{D,d} \rightarrow I^n$, as $d \rightarrow \infty$. 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 counter-example. Kent's example is a mixture of normals, which is $N_d(0, I_d)$ with probability 1/2 and $N_d(0, 10I_d)$ also with probability 1/2. This example satisfies conditions (i), (ii), (iv) and (iv). 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 (iii) is weaker than Gaussianity, assuming only a set of underlying independent entries, \mathbf{Z}_d^+ . We restate the theorem as follows.

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

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

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

COROLLARY 2. 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

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

Thus these n^+ data vectors form a regular $(n^+ - 1)$ -simplex in \mathbb{R}^d .

4.1.2 Geometric Representations for Two Samples

The *n*⁻-point sample $\mathcal{X}^{-}(d) = \{\mathbf{x}_{1}^{-}(d), \mathbf{x}_{2}^{-}(d), \dots, \mathbf{x}_{n^{-}}^{-}(d)\}$ is defined similarly to $\mathcal{X}^{+}(d)$. In particular, 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,

$$\frac{1}{d\tau_d^2} \|\mathbf{x}_k^- - \mathbf{x}_l^-\|^2 \to 2, \quad \text{as} \quad d \to \infty.$$
(10)

Now we generalize the two classes to allow different means. We assume that the squared distance between the population means, rescaled by 1/d, is a constant μ^2 ,

$$\frac{1}{d} \|E(\mathbf{x}^+) - E(\mathbf{x}^-)\|^2 \to \mu^2.$$
(11)

For convenience, we assume that the limiting average eigenvalues exist,

$$\sigma_d^2 \to \sigma^2 \text{ and } \tau_d^2 \to \tau^2 \text{ as } d \to \infty.$$
 (12)

THEOREM 3. Assume two independent data samples $\mathcal{X}^+(d)$ and $\mathcal{X}^-(d)$ satisfy Assumption 1, (11) and (12). 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 $l^2 := \sigma^2 + \tau^2 + \mu^2$, i.e.,

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

Theorem 3 says that, if both samples satisfy Assumption 1, then the pairwise rescaled distance between all pairs of data vectors from the two samples is approximately constant. Theorem 3 gives the interclass distances in the d-limit, while Corollary 2 and (10) 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 4.1.3 and Section

4.1.4, there is a natural separating hyperplane, that will give good classification, i.e., good generalization ability.

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 4.1.3 and the generalization ability is weak as discussed in Section 4.1.4.

4.1.3 Asymptotic Properties of the wDWD Intercept

In this section, we illustrate the asymptotic properties of the wDWD intercept in the HDLSS data settings. Let O^+ be the centroid of the $(n^+ - 1)$ -simplex $\mathcal{X}^+(d)$ and O^- the centroid of the $(n^- - 1)$ -simplex $\mathcal{X}^-(d)$. As noted in Hall et al. (2005), an important corollary of Corollary 2 and Theorem 3 is,

COROLLARY 4. In the d-asymptotic limit, the DWD hyperplane is orthogonal to the line O^+O^- joining the two centroids.

Let P be any point on the interval O^+O^- . In Figure 6, let α and β be the distances from P to the centroids. P lies on the weighted DWD hyperplane only when

$$\frac{\alpha}{\beta} = \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}.$$
(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 it will be more likely that a new data point will be classified to \mathcal{X}^+ . Theorem 5 shows the conditions under which a future data point is always correctly classified or misclassified.



Figure 6: Simplex centroids O^+ , O^- and the candidate DWD cut-off point P

THEOREM 5. 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,
 - If μ² > (n⁻w₋/n⁺w₊)^{1/2}σ²/n⁺ τ²/n⁻, then Pr(X₀⁺ is correctly classified by weighted DWD) → 1, as d→∞.
 If μ² < (n⁻w₋/n⁺w₊)^{1/2}σ²/n⁺ - τ²/n⁻, then Pr(X₀⁺ is wrongly classified by weighted DWD) → 1, as d→∞.
- For a new data point X₀⁻ from the X⁻-population, for any μ > 0,
 Pr(X₀⁻ is correctly classified by weighted DWD) → 1, as d → ∞.

An intuitive interpretation of Theorem 5 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 lead to better accuracy in general. When one class has a smaller intraclass variance or a larger sample size, standard DWD will give a more accurate classification rule. This comes at a cost of worse classification performance for the other class. Weighted DWD helps to offset the effect of unbalanced sample size to some extent.

Theorem 5 is the *weighted* extension to Theorem 3 in Hall et al. (2005). Compared to its original version, Theorem 5 extends DWD by the introduction of w_+ and $w_$ into the assumptions. 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 $\frac{w^+}{w^-} = \frac{n^-}{n^+}$ under MWGE. In Theorem 5, the main condition in Hall et al. (2005), $\sigma^2/n_+^{\frac{3}{2}} \ge \tau^2/n_-^{\frac{3}{2}}$, is relaxed to $\sigma^2/n^+ \ge \tau^2/n^-$. This condition is more easily satisfied so that, as shown in Theorem 5, one can classify a new data point from \mathcal{X}^- correctly by weighted DWD in contrast to standard 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 standard DWD in the asymptotic setting of Theorem 5, misclassifying some future points is unavoidable, 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 weighted DWD, we can adaptively choose the weights to adjust those relevant quantities, which can 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 5 can be avoided as much as possible.

4.1.4 Asymptotic Properties of the wDWD Direction

Theorem 5 gives a sufficient condition under which new data are correctly classified. 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. When the noise level is not negligible with respect to the signal (the interclass distance μ^2), Theorem 5 does not indicate the performance of wDWD. Instead, in this case, the relationship between the wDWD direction (the vector orthogonal to the separating hyperplane) and the direction of the line joining the two population means is more useful. If the angle between the above two directions is close to 0, the classification can be generalizable, in the sense of performing well for new data.

THEOREM 6. Assume that $\mathcal{X}^+(d)$ and $\mathcal{X}^-(d)$ satisfy Assumption 1. As $d \to \infty$, with probability converging to 1, the angle between the direction joining the two population means and the direction joining the centroids of the two simplices becomes $\theta = \cos^{-1} \left(\frac{\mu^2}{\mu^2 + \sigma^2/n^+ + \tau^2/n^-} \right)^{\frac{1}{2}}$.

Recall from Corollary 4, the weighted DWD direction coincides with the direction which joins the two centroids *d*-asymptotically. The asymptotic property of the angle θ between the wDWD direction and the optimal linear classification direction is then implied by Theorem 6. In particular,

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

in the sense that $\lim_{\gamma\to 0} \theta = 90^{\circ}$ and $\lim_{\gamma\to\infty} \theta = 0^{\circ}$ for $\gamma = \mu^2/(\frac{\sigma^2}{n^+} + \frac{\tau^2}{n^-})$. Theorem 6 and (14) imply that wDWD tends to give the optimal linear classification direction when the signal level μ^2 is much higher than the noise levels σ^2 and τ^2 , and on the other hand tends to give a direction which is orthogonal to the desired direction, i.e., is *strongly inconsistent*, when the noise is significantly greater than the signal. The second implication of Theorem 6 is that the angle goes to 0 if n^+ and $n^- \to \infty$, giving another notion of consistency of wDWD from the *d*-asymptotic point of view.

4.2 Simulation Confirmation

In this section, we verify the asymptotic results for weighted DWD by simulations. To verify Theorem 1, Corollary 2 and Theorem 3, which provide the interclass and intraclass pairwise distances, in Section 4.2.1, we calculate the corresponding distances for the high-dimensional simulated example discussed in Section 3.1.1. To verify Theorem 5 and Theorem 6, we perform a new simulation study in Section 4.2.2.

4.2.1 Pairwise Distances

We calculate the pairwise squared distances (scaled by d^{-1}) within each class and between classes for the constant signal simulation described in Section 3.1.1. Table 7 shows the summary statistics. In Table 7, note that all 3 of the mean rescaled squared distances fall reasonably close to the theoretical predictions. Moreover, the small standard deviation of the observed distance is consistent with Theorem 1 and Theorem 3, which imply that the distance should be constant in the large *d*-limit.

	# of pairs	mean	S.D.	theoretical	formula
within positive class	72010	1.1241	0.0489	1.1250	$2\sigma^2$
within negative class	191890	1.1242	0.0491	1.1250	$2\tau^2$
between classes	235600	1.1339	0.0491	1.1340	$\sigma^2 + \tau^2 + \mu^2$

Table 7: Summary statistics for the rescaled pairwise squared distances. The standard deviation of the distance is small relative to the mean.

4.2.2 DWD Classification Performance

To verify Theorem 5 and Theorem 6, we consider three simulated examples similar to the constant signal setting in Section 3.1.1. Here we fix the same noise level ($\sigma^2 = \tau^2 = 1$) and the sample sizes ($n^+ = 60$, $n^- = 150$), but assign different signal levels (μ^2) over the three examples. With the assumption of equal costs and equal class proportions, the optimal weights from Table 3 are $w_+ = \frac{1}{n^+}$ and $w_- = \frac{1}{n^-}$. Standard DWD is a special case of weighted DWD with $w_+ = w_- = 1$. Theorem 5 gives a threshold for μ^2 ,

$$(n^{-}w_{-}/n^{+}w_{+})^{\frac{1}{2}}\sigma^{2}/n^{+} - \tau^{2}/n^{-}.$$
(15)

According to the theorem, standard/weighted DWD correctly classifies X_0^+ with probability 1 if μ^2 is greater than the threshold. Here, the value of (15) for standard DWD is $(n^-/n^+)^{\frac{1}{2}}\sigma^2/n^+ - \tau^2/n^- = 0.020$, and that for weighted DWD it is $\sigma^2/n^+ - \tau^2/n^- = 0.010$. We explore the possible cases, by choosing

- $\mu^2 = 0.005$, where neither correct classification probability takes to 1;
- $\mu^2 = 0.011$, where only the wDWD correct classification probability takes to 1;
- $\mu^2 = 0.059$, where both wDWD and stdDWD correct classification probabilities take to 1.

	Case 1 weak $\mu^2 = 0.005 < 0.01$	Case 2 intermediate $0.01 < \mu^2 = 0.011 < 0.02$	Case 3 strong $\mu^2 = 0.059 > 0.02$	
	Class + Class -	Class + Class -	Class + Class -	
error wDWD	$43.91(1.714)\ 33.76(1.674)$	$29.41(1.368) \ 25.3(1.244)$	$2.4(0.148) \ 0.71(0.078)$	
error stdDWD	78.04(0.381) $8.63(0.233)$	$67.06(0.432) \ 4.67(0.148)$	$13.17(0.291) \ 0.05(0.015)$	
theoretical angle	65.16	55.43	32.15	
obs. angle wDWD	65.41(0.208)	55.59(0.189)	32.56(0.139)	
obs. angle stdDWD	66.86(0.193)	57.5(0.189)	34.06(0.127)	

Table 8: Simulation results for theorem verification: The top rows investigate Theorem 5; they display the average misclassification errors for both classes over 100 simulations and the standard error (in parentheses). The bottom rows validate Theorem 6, by showing that the theoretical angle between the DWD direction and the optimal classification direction given by the theorem, and the average observed angles for both wDWD and stdDWD together with the standard error (in parentheses).

In Table 8, note that when the signal is weak enough ($\mu^2 = 0.005$), both weighted and standard DWD fail to classify future data vectors from the \mathcal{X}^+ population. However, when the signal is strong enough ($\mu^2 = 0.059$), both methods succeed. If the data have intermediate signal strength ($\mu^2 = 0.011$), then weighted DWD works reasonably well (error< 30%) while the standard DWD does not (error> 60%). These observations are consistent with Theorem 5. Secondly, we find that the observed angles in the simulation for both weighted and standard DWD are in line with the theoretical angles based on the *d*-asymptotic results given by Theorem 6. Note that the angle between the optimal direction and the weighted DWD direction will often be closer to the theoretical angle (from Theorem 6), than that of the standard DWD.

4.3 Fisher Consistency of DWD

This section studies Fisher consistency of weighted DWD. As noted in Bartlett et al. (2006), many of the classification algorithms, developed in the machine learning literature, can be viewed as minimum contrast methods that minimize a convex surrogate of the 0-1 loss function. The weighted DWD (2) minimizes a surrogate of the corresponding weighted 0-1 loss function, $W(-1)I[y = -1]I[\phi(\mathbf{x}) = +1] + W(+1)I[y =$ $+1]I[\phi(\mathbf{x}) = -1]$. We first demonstrate the convex surrogate loss function for DWD (Section 2.2). This is similar to the Hinge loss function for SVM (Wahba 1999) through an equivalent formulation of the DWD optimization. A binary classifier with loss $V(yf(\mathbf{x}))$ is Fisher consistent if the minimizer of $E[W(Y_s)V(Y_Sf(\mathbf{X}_s))]$ has the same sign as $p_s(\mathbf{x}) - \frac{W(-1)}{W(-1)+W(+1)}$. Liu (2007) studied Fisher consistency for multicategorical SVM and its various extensions. To our knowledge, Fisher consistency of DWD has not been studied.

4.3.1 Equivalent Formulation

For each i = 1, ..., n, we define $f_i = f(\mathbf{x}_i | \boldsymbol{\omega}, b) = \mathbf{x}'_i \boldsymbol{\omega} + b$. The weighted DWD optimization problem (2) can be shown to be equivalent to the following problem.

$$\min_{\{\boldsymbol{\omega}, b: \; \boldsymbol{\omega}' \boldsymbol{\omega} \le 1\}} \min_{\boldsymbol{\xi} \ge 0} \sum_{i=1}^{n} W(y_i) \Big(\frac{1}{y_i f_i + \xi_i} + C\xi_i \Big). \tag{16}$$

It can be shown that the optimal solution for the inside optimization part of (16) is given by $\boldsymbol{\xi}^* = (\xi_1^*, \cdots, \xi_n^*)^T$, 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 the DWD optimization problem amounts to

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

If we define the *DWD loss function* as

$$V(yf) = \begin{cases} 2\sqrt{C} - C \cdot yf & \text{if } yf \leq \frac{1}{\sqrt{C}} \\ \frac{1}{yf} & \text{otherwise,} \end{cases}$$
(17)

then the weighted DWD optimization is $\min_{\boldsymbol{\omega}, b} \sum_{i=1}^{n} W(y_i) V(y_i f_i(\boldsymbol{\omega}, b))$, s.t. $\boldsymbol{\omega}' \boldsymbol{\omega} \leq 1$. This representation provides some insights into DWD as a modification of the Hinge loss of SVM, $H(yf) = (1 - yf)_+$. Actually, the first expression for the DWD loss is similar to the Hinge loss, while the second expression $\frac{1}{yf}$ is positive, rather than 0 for the Hinge loss when yf > 1.

4.3.2 Fisher Consistency

For any classification function f, the expected DWD loss, i.e., the risk, is $R(f) = E[W(Y_s)V(Y_sf(\mathbf{X}_s))]$. Fisher consistency of the classifier f can be proved by showing that the sign of the global minimizer of the unconditional risk arg min_f R(f), is equal to the Bayes optimal decision rule ϕ^* given in (4). Theorem 7 proves this relationship and thus shows Fisher consistency of weighted DWD under the OM criterion.

THEOREM 7. Let f^* be the global minimizer of $E[W(Y_s)V(Y_sf(\mathbf{X}_s))]$, where $V(\cdot)$ is the DWD loss function given in (17). Then $sign[f^*(\mathbf{x})] = \phi^*(\mathbf{x})$, where $\phi^*(\mathbf{x})$ is the Bayes decision rule under the OM criterion given in (4), or equivalently, $sign[f^*(\mathbf{x})] =$ $sign[p_s(\mathbf{x}) - \frac{W(-1)}{W(+1)+W(-1)}]$. Similarly, under the MWGE criterion, with the weighting scheme $W(\cdot)$ given by Table 3, weighted DWD can also be shown to be Fisher consistent.

5 Conclusion

In this article, we have proposed weighted DWD to improve standard DWD for unbalanced data and various nonstandard situations. We have made the following contributions. First of all, we have provided the optimal weighting schemes for several nonstandard situations, using one of the two criteria, OM rate and MWGE. Secondly, we represent data sets from two classes geometrically in HDLSS settings. Thirdly, we develop the HDLSS asymptotic properties of weighted DWD. Lastly, we show Fisher Consistency for wDWD. Our numerical studies demonstrate the effectiveness of weighted DWD and verify the asymptotic results.

The results on the tuning parameter C from our simulations suggest that the recommendation for the tuning parameter $C = 100/(dt)^2$ proposed by Marron et al. (2007), which was originally designed for balanced data, also works well in unbalanced and nonstandard situations as long as we use weighted DWD instead of standard DWD. Thus their recommendation of tuning parameter C can be used for weighted DWD as a simple alternative of cross validation.

The simulation results show that in the sparse signal setting, our current version of weighted DWD does not work as well as some sparse methods, for example L_1 SVM. One possible future research direction is to study weighted DWD with built-in sparse penalty for variable selection.

Appendix

For Theorem 1, Corollary 2 and Theorem 5, we only outline the main steps of the proofs. Readers can refer to Qiao et al. (2008) for technical details.

Proof of Theorem 1:

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 (8), $\frac{1}{\sigma_d^2}S_{D,d}^+ = \frac{1}{\Sigma_{i=1}^d\lambda_{i,d}^+}\Sigma_{i=1}^d(\lambda_{i,d}^+W_{i,d}^+) = \Sigma_{i=1}^d(\tilde{\lambda}_{i,d}^+W_{i,d}^+)$. 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 $(i = 1, \dots, d)$ 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 $(i = 1, \dots, d)$, with mean 0 and unit variance.

Chebyshev's inequality is then used twice (one for the diagonal elements, one for the off-diagonal elements) to show that 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 the multivariate Gaussian distribution, so does \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 2

Let $\mathbf{x}_j^+ = (\mathbf{x}_{1j}^+, \cdots, \mathbf{x}_{dj}^+)^T$, $j = 1, \cdots, n^+$, be the *j*th 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 *j*th 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 $\frac{1}{d\sigma_d^2} \|\mathbf{x}_k^+ - \mathbf{x}_l^+\|^2 = \frac{1}{d\sigma_d^2} \sum_{i=1}^d (\mathbf{x}_{ik}^+ - \mathbf{x}_{il}^+)^2 = \frac{1}{d\sigma_d^2} \sum_{i=1}^d (\mathbf{x}_{ik}^+)^2 + \frac{1}{d\sigma_d^2} \sum_{i=1}^d (\mathbf{x}_{il}^+)^2 - \frac{2}{d\sigma_d^2} \sum_{i=1}^d \mathbf{x}_{ik} \mathbf{x}_{il}$. The first and second terms on the right and 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 1. 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 8. 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 by $U = [u_{ij}]_{i,j=1,\dots,d}$ 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 $d \to \infty$.

Note that sum of squared entries in each column and row of U is 1. Lemma 8 can be proved using the Cauchy-Schwarz inequality.

Proof of Theorem 3

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

$$\|\mathbf{x}_{k}^{+} - \mathbf{x}_{l}^{-}\|^{2} = \Sigma_{i=1}^{d} \left\{ [\mathbf{x}_{ik}^{+} - E(\mathbf{x}_{i\cdot}^{+})] - [\mathbf{x}_{il}^{-} - E(\mathbf{x}_{i\cdot}^{-})] + [E(\mathbf{x}_{i\cdot}^{+}) - E(\mathbf{x}_{i\cdot}^{-})] \right\}^{2}$$
(18)

$$= \Sigma_{i=1}^{d} (\dot{\mathbf{x}}_{ik}^{+})^{2} + \Sigma_{i=1}^{d} (\dot{\mathbf{x}}_{il}^{-})^{2} - 2\Sigma_{i=1}^{d} (\dot{\mathbf{x}}_{ik}^{+}) (\dot{\mathbf{x}}_{il}^{-})$$
(19)

$$+\Sigma_{i=1}^{d} [E(\mathbf{x}_{i}^{+}) - E(\mathbf{x}_{i}^{-})]^{2} + 2\Sigma_{i=1}^{d} [E(\mathbf{x}_{i}^{+}) - E(\mathbf{x}_{i}^{-})][\dot{\mathbf{x}}_{ik}^{+} - \dot{\mathbf{x}}_{il}^{-}]. (20)$$

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 *i*th entries on the *k*th and *l*th columns of the *de-meaned* data matrices $\dot{\mathbf{X}}^+$ and $\dot{\mathbf{X}}^-$.

The first two terms in (19), 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 1, both converge to 1 in probability as $d \to \infty$. Thus, for any $\varepsilon > 0$, $Pr(|\frac{1}{d}\Sigma_{i=1}^d(\dot{\mathbf{x}}_{ik}^+)^2 - \sigma^2| \ge \varepsilon) \to 0$, as $d \to \infty$ and $Pr(|\frac{1}{d}\Sigma_{i=1}^d(\dot{\mathbf{x}}_{il}^-)^2 - \tau^2| \ge \varepsilon) \to 0$, 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 de-meaned 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 $\tilde{\lambda}_{j,d}^{-} = \lambda_{j,d}^{-}/\Sigma_{j=1}^{d}\lambda_{j,d}^{-}$. $(d\sigma_d \tau_d)^{-1} \Sigma_{i=1}^d (\dot{\mathbf{x}}_{ik}^+) (\dot{\mathbf{x}}_{il}^-)$ becomes

$$(d\sigma_{d}\tau_{d})^{-1}[z_{1}^{+},\cdots,z_{d}^{+}](\Lambda^{+})^{\frac{1}{2}}V^{+T}V^{-}(\Lambda^{-})^{\frac{1}{2}}[z_{1}^{-},\cdots,z_{d}^{-}]^{T}$$

$$= (\Sigma_{i=1}^{d}\lambda_{i,d}^{+})^{-\frac{1}{2}}(\Sigma_{j=1}^{d}\lambda_{j,d}^{-})^{-\frac{1}{2}}\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_{s}^{+}z_{t}^{-}\sqrt{\lambda_{s,d}^{+}\lambda_{t,d}^{-}}$$

$$= \Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_{s}^{+}z_{t}^{-}\sqrt{\tilde{\lambda}_{s,d}^{+}\tilde{\lambda}_{t,d}^{-}}$$

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,

$$Pr[|\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_{s}^{+}z_{t}^{-}\sqrt{\tilde{\lambda}_{s,d}^{+}\tilde{\lambda}_{t,d}^{-}}| \ge \varepsilon]$$

$$\leq \varepsilon^{-2}E(\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_{s}^{+}z_{t}^{-}\sqrt{\tilde{\lambda}_{s,d}^{+}\tilde{\lambda}_{t,d}^{-}})^{2}$$

$$= \varepsilon^{-2}\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}^{2}\tilde{\lambda}_{s,d}^{+}\tilde{\lambda}_{t,d}^{-}$$

Since U is the product of two orthogonal matrix $U = V^{+T}V^{-}$, U is itself orthogonal. nal. The relative eigenvalues satisfy the condition in Lemma 8. Thus by Lemma 8, $Pr[|\Sigma_{s=1}^{d}\Sigma_{t=1}^{d}u_{s,t}z_{s}^{+}z_{t}^{-}\sqrt{\tilde{\lambda}_{s,d}^{+}\tilde{\lambda}_{t,d}^{-}}| \geq \varepsilon] \rightarrow 0$, 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 means, which is defined as $d\mu^2$.

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

$$\begin{aligned} ⪻(|\frac{1}{d^{\frac{1}{2}}}\Sigma_{i=1}^{d}\delta_{i}\Sigma_{s=1}^{d}v_{i,s}^{+}\sqrt{\tilde{\lambda}_{s,d}^{+}}z_{s}^{+}| > \varepsilon) \\ &\leq \quad \varepsilon^{-2}E(\frac{1}{d^{\frac{1}{2}}}\Sigma_{i=1}^{d}\delta_{i}\Sigma_{s=1}^{d}v_{i,s}^{+}\sqrt{\tilde{\lambda}_{s,d}^{+}}z_{s}^{+})^{2} \quad = \varepsilon^{-2}\frac{1}{d}E(\Sigma_{s=1}^{d}\Sigma_{i=1}^{d}\delta_{i}v_{i,s}^{+}\sqrt{\tilde{\lambda}_{s,d}^{+}}z_{s}^{+})^{2} \\ &= \quad \varepsilon^{-2}\frac{1}{d}\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_{i}v_{i,s}^{+})^{2}\tilde{\lambda}_{s,d}^{+}E(z_{s}^{+})^{2} \quad = \varepsilon^{-2}\frac{1}{d}\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_{i}v_{i,s}^{+})^{2}\tilde{\lambda}_{s,d}^{+} \\ &\leq \quad \varepsilon^{-2}\frac{1}{d}\Sigma_{s=1}^{d}(\Sigma_{i=1}^{d}\delta_{i}v_{i,s}^{+})^{2}\max_{i}(\tilde{\lambda}_{i,d}^{+}) \quad = \varepsilon^{-2}\mu^{2}\max_{i}(\tilde{\lambda}_{i,d}^{+}) \to 0, \quad \text{as} \quad d \to \infty. \end{aligned}$$

Note that $\sum_{s=1}^{d} (\sum_{i=1}^{d} \delta_i v_{i,s}^+)^2 = \sum_{i=1}^{d} \delta_i^2 = d\mu^2$ because V^+ is an orthogonal matrix, which keeps the norm of δ after transformation. Hence the first component $\sum_{i=1}^{d} [E(\mathbf{x}_{i}^+) - \mathbf{x}_{i}^+)]$

 $E(\mathbf{x}_{i\cdot}^{-})]\dot{\mathbf{x}}_{ik}^{+}$, rescaled by d^{-1} , converges to 0 in probability as $d \to \infty$. And so does the second component $\sum_{i=1}^{d} [E(\mathbf{x}_{i\cdot}^{+}) - E(\mathbf{x}_{i\cdot}^{-})]\dot{\mathbf{x}}_{il}^{-}$.

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

Proof of Theorem 5:

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 . It was shown by a series of geometric calculations in Hall et al. (2005) that $\frac{\alpha}{\beta} = \frac{\sigma^2/n^+}{\mu^2 + \tau^2/n^-}$.

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 $\frac{\sigma^2/n^+}{\mu^2 + \tau^2/n^-} < \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}$. It will be wrongly classified as \mathcal{X}^- if $\frac{\sigma^2/n^+}{\mu^2 + \tau^2/n^-} > \left(\frac{w_+ n^+}{w_- n^-}\right)^{1/2}$.

The first and second parts of Theorem 5 follows from the two inequalities above immediately. 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 5 simply combines the analysis above.

Proof of Theorem 6:

Denote the centroids of the $(n^+ - 1)$ -simplex from \mathcal{X}^+ as $O^{n^+}_+$ and the $(n^- - 1)$ simplex from X^- as $O^{n^-}_-$. Also denote the population means of \mathcal{X}^+ and \mathcal{X}^- as O^{∞}_+ and $O_{-}^{\infty} \text{ respectively. In the large } d\text{-limit, the expected squared distance, rescaled by } d^{-1},$ between $O_{+}^{n^+}$ and $O_{-}^{n^-}$ is $\mu^2 + \sigma^2/n^+ + \tau^2/n^-$. If we consider k more data vectors from \mathcal{X}^+ , the expected squared distance, rescaled by d^{-1} , between the centroids $O_{+}^{(n^++k)}$, of the new $(n^+ + k - 1)$ -simplex, and the centroid $O_{-}^{n^-}$, of the $(n^- - 1)$ -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^+-1) -simplex $\{\sqrt{d}(\underbrace{1, 0, \cdots, 0}_{n^+, k}, \underbrace{0, 0, \cdots, 0}_{n^++k}, \underbrace{\sqrt{d}(\underbrace{1, 0, \cdots, 0}_{n^++k}, \underbrace{0, 0, \cdots, 0}_{n^++k}, \underbrace{1, 0, \cdots, 0}_{n^++k})^T$ 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 \mathcal{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 $(n^+ - 1)$ -simplex $\mathcal{X}^+(d)$ and the $(n^- - 1)$ -simplex $\mathcal{X}^-(d)$ becomes $\theta = \cos^{-1}\left(\frac{\mu^2}{\mu^2 + \sigma^2/n^+ + \tau^2/n^-}\right)^{\frac{1}{2}}$.

Proof of Theorem 7:

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

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

Here the DWD loss $V(\cdot)$ is defined in (17). 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) = 0

 $p_s W(+1) 1 V'(f) - (1 - p_s) W(-1) V'(-f)$. Straightforward computations give

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

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 directly solving the equation R'(f) = 0, we get f^* , the minimizer of R(f) as

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

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

References

- Ahn, J., Marron., J. S., Muller, K. M. and Chi, Y. (2007), "The High-dimension, Lowsample-size Geometric Representation Holds Under Mild Conditions" *Biometrika*, 94, 3, 760-766.
- Alizadeh, F. and Goldfarb, D. (2003), "Second-Order Cone Programming", Mathematical Programming, 95, 3-51.
- Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006), "Convexity, Classification, and Risk Bounds", Journal of the American Statistical Association, 101, 473, 138-156.
- 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 Pro-

filing Reveals Distinct Adenocarcinoma Subclasses", *Proc Natl Acad Sci U S A*, 98(24):13790-5.

- Chen, P. H., Lin, C. J., and Schölkopf, B. (2005), "A Tutorial on *v*-Support Vector Machines", Applied Stochastic Models in Business and Industry, 21(2), 111-136.
- 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.
- Dudoit, S., Fridlyand, J. and Speed T. P. (2002), "Comparison of discrimination methods for the classification of tumors using gene expression data", Journal of the American Statistical Association, 97, 77-87.
- Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American Statistical Association, 96, 456, 1348-1360.
- Fung, G. M., Mangasarian, O. L. (2004), "A feature selection Newton method for support vector machine classification", *Computational Optimization and Applications*, 28, 185-202.
- Ge, N. and Simpson, D. G. (1998), "Correlation and High-Dimensional Consistency in Pattern Recognition", Journal of the American Statistical Association, 93, 443, 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, 427 - 444.
- Jung, S. K. and Marron, J. S. (2008), "PCA Consistency in High Dimension, Low Sample Size Context", The Annals of Statistics, to appear.
- Le Cessie, S. and Van Houwelingen, J. (1992), "Ridge estimators in logistic regression", Applied Statistics, 41, 191-201.

- Lee, A. and Silvapulle, M. (1988), "Ridge estimation in logistic regression", Communications in Statistics, Simulation and Computation 17, 1231-1257.
- Lin, Y., Lee, Y., and Wahba, G. (2002), "Support Vector Machine for Classification in Nonstandard Situation", *Machine Learning* 46, 191-202.
- Liu, Y. (2007). "Fisher consistency of multicategory support vector machines", *Eleventh* International Conference on Artificial Intelligence and Statistics, 289-296.
- Liu, Y., Hayes, D. N., Nobel, A. and Marron, J. S. (2008), "Statistical Significance of Clustering for High Dimension Low Sample Size Data", Journal of the American Statistical Association, 103, 483, 1281-1293.
- Lokhorst, J. (1999), "The lasso and generalised linear models", *Technical Report*, University of Adelaide, Adelaide.
- Marron, J. S., Todd, M. and Ahn, J. (2007), "Distance Weighted Discrimination", Journal of the American Statistical Association, 102, 480, 1267-1271(5).
- Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J. and Marron, J. S. (2008), "Asymptotic Properties of Distance-Weighted Discrimination", *manuscript*.
- Qiao, X. and Liu, Y. (2009), "Adaptive Weighted Learning for Unbalanced Multicategory Classification", *Biometrics*, 65, 1, 159-168.
- Schölkopf, B. and A.J. Smola (2002), Learning with Kernels, MIT Press, Cambridge.
- Shevade, S. and Keerthi, S. (2003), "A simple and efficient algorithm for gene selection using sparse logistic regression", *Bioinformatics*, 19, 2246-253.
- Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society* Series B, 58, 267-288.
- Vapnik, V. N. (1995), The Nature of Statistical Learning Theory, Springer-Verlag, Berlin
- Wahba, G. (1999), "Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV", in Advances in Kernel Methods Support Vector Learning, eds., Schölkopf, B., Burges, C. and Smola, A., MIT Press, Cambridge, 69-88.