Distance Weighted Discrimination

J. S. Marron, Michael J. Todd and Jeongyoun Ahn

Abstract

High Dimension Low Sample Size statistical analysis is becoming increasingly important in a wide range of applied contexts. In such situations, it is seen that the popular Support Vector Machine suffers from "data piling" at the margin, which can diminish generalizability. This leads naturally to the development of Distance Weighted Discrimination, which is based on Second Order Cone Programming, a modern computationally intensive optimization method.

1 Introduction

An area of emerging importance in statistics is the analysis of High Dimension Low Sample Size (HDLSS) data. This area can be viewed as a subset of multivariate analysis, where the dimension d of the data vectors is larger than the sample size n. In this paper, the focus is on two-class discrimination. A clever and powerful discrimination method is the support vector machine (SVM) (Vapnik 1995; Cristianini

and Shawe-Taylor 2000). The first contribution of the present paper is a novel view of the performance of SVM, in HDLSS settings, via projecting the data onto the normal vector of the separating hyperplane. This view reveals substantial *data piling* of SVM, which means that many of these projections are identical.

The discussion below suggests that data piling may adversely affect the generalization performance of SVM in some HDLSS situations. The major contribution of this paper is a new discrimination method, called Distance Weighted Discrimination (DWD), which avoids data piling, and is seen in some examples to give the anticipated improved generalizability. The DWD uses interior-point methods for so-called second-order cone programming (SOCP) problems (Alizadeh and Goldfarb 2003).

The two-class discrimination problem begins with two sets (classes) of d-dimensional training data vectors. The goal of discrimination is to find a rule for assigning the labels of $+1$ or -1 to new data vectors, depending on whether the vectors are "more like Class $+1$" or are "more like Class -1."

For simplicity only linear discrimination methods are considered here. (Extensions to the nonlinear case are discussed in Section 2.) In particular, there is a direction vector $\mathbf{w}$ and a threshold β, so that the new data vector $\mathbf{x}$ is assigned to the Class $+1$ when $\mathbf{x}'\mathbf{w} + \beta \geqslant 0$. This corresponds to separation of the d-dimensional data space into two regions by a hyperplane, with normal vector $\mathbf{w}$, whose position is determined by β. Data piling occurs when many data points have identical projections in that direction, i.e., the data pile up on top of each other. More relevant to the discrimination problem, it means that at least some of the data points pile at two different common points, one for each class.

The key idea behind SVM is to find $\mathbf{w}$ and β to keep the data in the same class all on the same side of, and also as far as possible from, the separating hyperplane. This is quantified using a maximin optimization formulation, focussing on only the

2

data points that are closest to the separating hyperplane, called *support vectors*. For SVM, data piling is common in HDLSS contexts because the support vectors (which tend to be very numerous in higher dimensions) all pile up at the boundaries of the margin when projected in this direction.

The simulated data in Figure 1 have dimension $d = 39$, with $n_+ = 20$ Class $+1$ samples represented as red plus signs, and $n_- = 20$ Class -1 samples represented as blue circles. The data were drawn from two Gaussian distributions with zero mean and identity covariance, whose mean in the first dimension is shifted to $+2.2$ (-2.2) for Class $+1$ (-1). Because the true difference lies only in the first coordinate direction, this is the normal vector of the theoretically Bayes optimal rule. Discrimination methods whose normal vector lies close to this direction should have good generalization properties.

Figure 1(a) shows one-dimensional projections of the data to the optimal normal vector. The data are represented as a "jitter plot," with the horizontal coordinate representing the projection, and with a random vertical coordinate used for visual separation of the points. Also kernel density estimates are included in order to give another indication of the structure of these univariate populations. As expected, it reveals two Gaussian populations, with respective means ± 2.2.

Figure 1(b) shows a different one-dimensional projection, this time based on the *maximal data piling* direction. It is seen in Ahn and Marron (2004) that for general HDLSS discrimination problems, there are a number of direction vectors which have complete data piling. Out of all of these, there is a unique direction, which maximizes the separation of the two piling points, called the maximal data piling (MDP) direction. The figure shows that indeed the data completely line up orthogonally to the MDP direction.

Data piling is usually not a useful property for a discrimination rule, because it
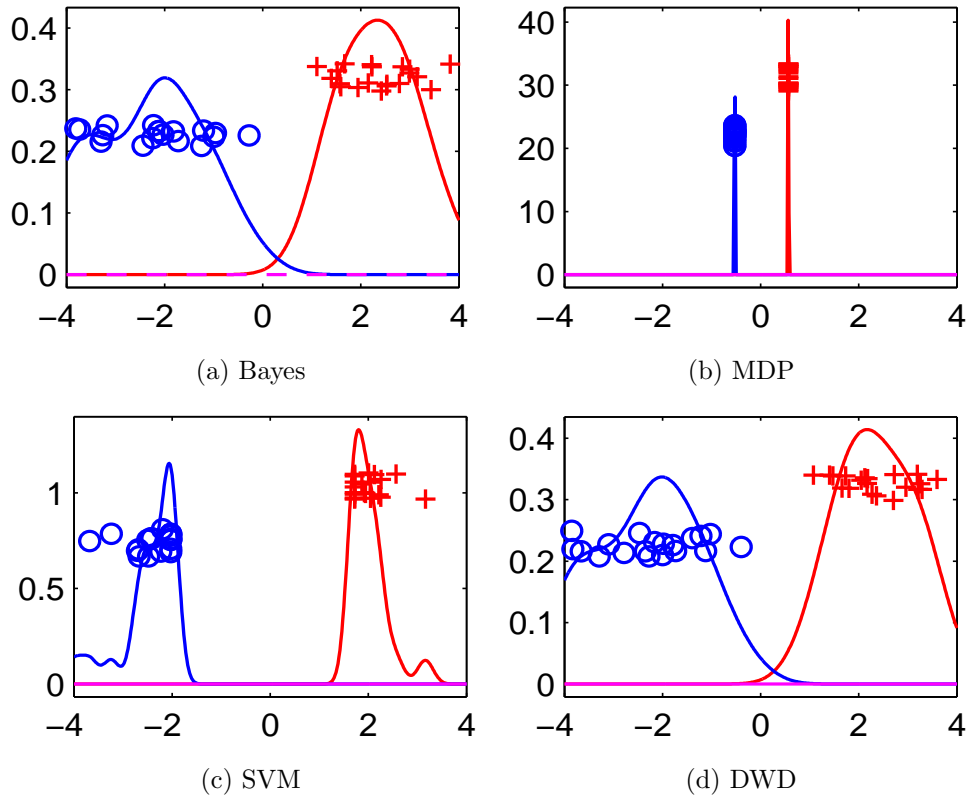
Figure 1: Toy example, illustrating potential for data piling in HDLSS settings. One-dimensional projections for four discriminant directions. (x-axis shows projections, y-axis is density). Directions with data piling such as MDP and SVM suggest worse generalizability.

is driven only by very particular aspects of the realization of the training data at hand. In HDLSS settings with a reasonable amount of noise, a different realization of the data will have its own quite different quirks, and thus will result in a completely different MDP direction. In this sense, the maximal data piling direction will typically not have good generalizability properties. Another way of understanding the poor generalizability comes from noting that the angle between the Bayes optimal and the MDP direction is quite large, at $56.3°$.

The data piling properties of SVM are shown in Figure 1(c). The angle, now much smaller at $35.8°$, reflects improved generalizability. However there is a clear piling up

of data at the margin, which suggests that there is room for improvement. This piling indicates that the algorithm is doing some overfitting of spurious noise artifacts in the data. This is inevitable in HDLSS situations, because in higher dimensions there will be more support vectors (i.e., data points right on the margin). The data piling of SVM can be reduced to some extent with careful tuning; however, we observed in some real data examples that data piling appears no matter how it is tuned.

Room for improvement comes from allowing more of the data points to have a direct impact on $\mathbf{w}$. In Section 2 we propose the new Distance Weighted Discrimination (DWD) method. Its optimization replaces the margin based criterion of SVM by a different function of the distances, r_i, from the data to the separating hyperplane. A simple way of allowing these distances to influence $\mathbf{w}$ is to optimize the sum of the inverse distances. This gives high significance to those points that are close to the hyperplane, with little impact from points that are farther away: hence distance weighted.

Figure 1(d) shows the projection onto the DWD direction, which is much closer to the Bayes optimal, with the angle now down to 25.4°. The plot shows no data piling. The improved generalizability of DWD over SVM has been verified using an appropriate type of asymptotic mathematics in Hall, Marron and Neeman (2005). An additional advantage of the approximately Gaussian sub-population shapes is that DWD provides a natural way to perform microarray bias adjustment (where the two sub-populations are moved along this vector until they overlap to remove data biases), as proposed in Benito et al. (2004).

A precise formulation of the optimization in DWD and some computational issues are discussed in Section 2. Section 3 contains some simulation studies. The main lesson is that every discrimination rule has some setting where it is best. The main strength of DWD is that its performance is close to the best method in each simulation

setting. Similar overall performance of DWD is shown on some real data examples in Section 4. See Ahn (2007) for additional material including earlier original (much more detailed) drafts of this paper.

2   Formulation of Optimization Problems

Let us first set the notation to be used. The training data consists of n d-vectors $\mathbf{x}_i$ together with corresponding class indicators $y_i \in \{+1, -1\}$. We let $\mathbf{X}$ denote the $d \times n$ matrix whose columns are the $\mathbf{x}_i$'s, and $\mathbf{y}$ the n-vector of the y_i's. It is convenient to use $\mathbf{Y}$ for the $n \times n$ diagonal matrix with the components of $\mathbf{y}$ on its diagonal. Then, if we choose $\mathbf{w} \in \mathbb{R}^d$ as the normal vector for our hyperplane and $\beta \in \mathbb{R}$ to determine its position, the residual of the ith data point is $\bar{r}_i = y_i(\mathbf{x}_i'\mathbf{w} + \beta)$, or in matrix-vector notation $\bar{\mathbf{r}} = \mathbf{Y}(\mathbf{X}'\mathbf{w} + \beta\mathbf{e}) = \mathbf{YX}'\mathbf{w} + \beta\mathbf{y}$, where $\mathbf{e} \in \mathbb{R}^n$ denotes the vector of ones. We would like to choose $\mathbf{w}$ and β so that all $\bar{r}_i$ are positive and "reasonably large." Of course, the $\bar{r}_i$'s can be made as large as we wish by scaling $\mathbf{w}$ and β, so $\mathbf{w}$ is scaled to have unit norm so that the residuals measure the signed distances of the points from the hyperplane.

However, it may not be possible to separate the positive and negative data points linearly, so we allow a vector $\boldsymbol{\xi} \in \mathbb{R}^n_+$ of errors, to be suitably penalized, and define the perturbed residuals to be $\mathbf{r} = \mathbf{YX}'\mathbf{w} + \beta\mathbf{y} + \boldsymbol{\xi}$. When the data vector $\mathbf{x}_i$ lies on the proper side of the separating hyperplane and the penalization is not too small, $\xi_i = 0$, and thus $\bar{r}_i = r_i$.

The SVM chooses $\mathbf{w}$ and β to maximize the minimum r_i in some sense, while our DWD approach instead minimizes the sum of reciprocals of the r_i's augmented by a penalty term. Both methods involve a tuning parameter that controls the penalization of $\boldsymbol{\xi}$, whose choice is discussed below. Note that Dandurova, Yeganova and Falk (2001) also propose an optimization problem for discrimination that uses a

nonlinear function (the negative exponential) of all the residuals, but their problem is nonconvex; their paper contains no computational results.

We now describe how the optimization problem for our new approach is defined. We choose as our new criterion that the sum of the reciprocals of the residuals, perturbed by a penalized vector $\boldsymbol{\xi}$, be minimized: thus we have

$$\min_{\mathbf{r},\mathbf{w},\beta,\boldsymbol{\xi}} \quad \sum_i \frac{1}{r_i} + C\mathbf{e}'\boldsymbol{\xi}, \quad \mathbf{r} = \mathbf{YX}'\mathbf{w} + \beta\mathbf{y} + \boldsymbol{\xi} \geqslant 0, \ \|\mathbf{w}\|^2 \leqslant 1, \ \boldsymbol{\xi} \geqslant 0, \qquad (1)$$

where again C is a penalty parameter. (We have relaxed the condition that the norm of $\mathbf{w}$ be equal to 1 to a requirement that it be at most 1. This makes the problem convex, and if the data are strictly linearly separable and C is large enough, the optimal solution will have norm equal to 1.)

While the discussion here is mostly on linear discrimination methods, it is important to note that this actually entails a much larger class of discriminators, through "polynomial embedding" and "kernel embedding" ideas; we call this the nonlinear case. This idea goes back at least to Aizerman, Braverman and Rozoner (1964) and involves either enhancing (or perhaps replacing) the data values with additional functions of the data. Such functions could involve powers of the data, in the case of polynomial embedding, or radial or sigmoidal kernel functions of the data. The nonlinear case is treated in the extended versions of this paper, available at Ahn (2007).

In the initial version of this paper available at Ahn (2007), we show how, using additional variables and constraints, this problem can be reformulated as a second-order cone programming (SOCP) problem. This is a problem with a linear objective, linear constraints, and the requirement that various subvectors of the decision vector must lie in second-order cones of the form $S_{m+1} := \{(\zeta; \mathbf{u}) \in \mathbb{R}^{m+1} : \zeta \geqslant \|\mathbf{u}\|\}$. For $m = 0$, 1, and 2, this cone is the nonnegative real line, a (rotated) quadrant, and the

right cone with axis $(1; 0; 0)$ respectively.

SOCP problems have nice duals, and after some algebra, we obtain a simplified form of the dual as

$$\max_{\boldsymbol{\alpha}} \quad -\|\mathbf{X Y}\boldsymbol{\alpha}\| + 2\mathbf{e}'\sqrt{\boldsymbol{\alpha}}, \quad \mathbf{y}'\boldsymbol{\alpha} = 0, \ 0 \leqslant \boldsymbol{\alpha} \leqslant C\mathbf{e}. \tag{2}$$

(Here $\sqrt{\boldsymbol{\alpha}}$ denotes the vector whose components are the square roots of those of $\boldsymbol{\alpha}$.) This can be compared with the dual of the SVM problem, which is identical except for having objective function $-(1/2)\|\mathbf{X Y}\boldsymbol{\alpha}\|^2 + \mathbf{e}'\boldsymbol{\alpha}$. Note that both problems have optimal solutions.

The aforementioned extended manuscript describes the necessary and sufficient optimality conditions for these problems, how the dual problem can be viewed as minimizing the distance between points in the convex hulls of the Class $+1$ points and of the Class -1 points, but now divided by the square of the sum of the square roots of the convex weights, and how DWD can be extended to the nonlinear case using a kernel function.

Problem (1) has $2n + 1$ equations and $3n + d + 2$ variables, and so solving it can be expensive in the large-scale HDLSS case. Our extended manuscript also shows how a preprocessing step can be performed to reduce d to n. From the optimality conditions, we can see that, if all x_i's are scaled by a factor γ, the penalty parameter C should be scaled by γ^{-2}, while if each training point is replicated p times, then C remains the same. Hence a reasonable value for C is some large constant divided by a typical distance between $\mathbf{x}_i$'s squared. As a notion of typical distance, we suggest the median of the pairwise Euclidean distances between classes, $d_t = \text{median} \{\|\mathbf{x}_i - \mathbf{x}_{i'}\| : \mathbf{y}_i = +1, \mathbf{y}_{i'} = -1\}$. In most examples in this paper, we use $C = 100/d_t^2$ for DWD and use Gunn's recommendation of $C = 1000$ for SVM (Gunn 1997). We employed cross-validation in the real data examples in Section 4.

For the SOCP problems (1) and (2), there are efficient primal-dual interior-point methods (see, e.g., Tütüncü, Toh and Todd (2003)), but active-set methods are still under development and untested. Thus in the large-scale case, the computational cost seems rather greater than in the SVM case. A small number of iterations is required (theoretically $O(\sqrt{n}\ln(1/\epsilon))$ to attain accuracy ϵ, but usually much fewer), but each requires the formation of an $n \times n$ linear system at a cost of $O(n^2 \max\{n, d\})$ operations, and then the solution of the system at a cost of $O(n^3)$ operations. Here each iteration can be viewed as a Newton-like step for related barrier problems or perturbed optimality conditions. In the HDLSS case, with $d \gg n$, by using the dimension-reduction procedure described in the extended manuscript, we can do an initial preprocessing of the data at a cost of $O(n^2 d)$ operations, and then each iteration requires $O(n^3)$ operations.

3   Simulations

In this section, we compare DWD with SVM, mean difference (MD) (i.e., the centroid method) and the regularized logistic regression (RLR) method of le Cessie and van Houwelingen (1992) in various HDLSS settings. For the RLR tuning parameter, we use the equivalent value to that of SVM.

In the simulation study presented here, for each example, training data sets of size $n_+ = n_- = 25$ and testing data sets of size 200, of dimensions $d = 10, 40, 100, 400, 1600$ were generated. The dimensions are intended to cover from non-HDLSS to extreme-HDLSS settings. Each experiment was replicated 100 times. The graphics in Figure 2 summarize the mean of the misclassification rates of test data with 95% confidence intervals as error bars.

The first simulation setting is the same as the toy example shown in Figure 1, except for the dimensions. For this example, MD is the empirically optimal Bayes

(a) Gaussian

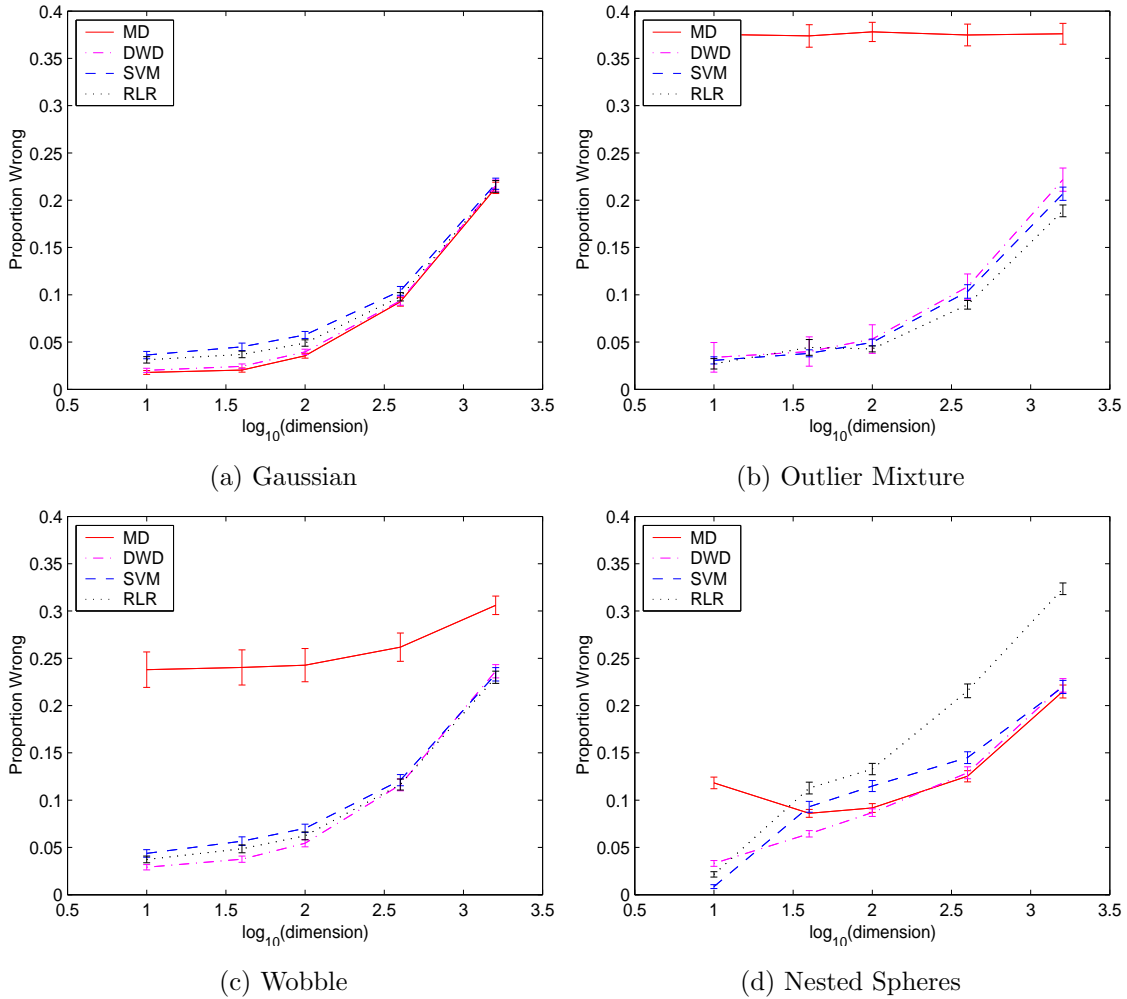(b) Outlier Mixture

(c) Wobble

(d) Nested Spheres

Figure 2: Summary of four simulations. DWD is always among the best.

rule, so the other methods are expected to have a worse error rate. In Figure 2(a), DWD shows very similar performance to MD, and the confidence interval indicates that the difference between the two methods is not statistically significant. Note that SVM has substantially worse error, due to the data piling effect.

The second simulation, for the so-called *outlier mixture* distribution, is a mixture distribution where 80% of the data are from the distributions from the previous example, and the remaining 20% are Gaussian with mean $+100$ (-100) in the first coordinate, $+500$ (-500) in the second coordinate for Class $+1$ (-1), and 0 in the other coordinates. Excellent discrimination for this problem is again provided by the

first coordinate axis direction, because the outliers are on the correct side. SVM is expected to have similar performance to the previous example, because the outliers will never be support vectors, as shown in Figure 2(b). RLR is the best among the four methods in this example and MD has very poor error rate because the sample means are dramatically impacted by the 20% outliers in the data. DWD nearly shares the good properties of the RLR because the outliers receive a very small weight.

Figure 2(c) shows an example where DWD is actually the best of these four methods. Here the data are from the *wobble distribution*, which is again a mixture, where again 80% of the data are from the shifted spherical Gaussian as in Figure 2(a), and the remaining 20% are chosen so that the first coordinate is replaced by $+0.1$ (-0.1), and just one randomly chosen coordinate is replaced by $+100$ (-100), for an observation from Class $+1$ (-1). That is, a few pairs of observations are chosen to violate the ideal margin, in ways that push directly on the support vectors. As in Figure 2(b), the few outliers have a serious and drastic effect on MD, giving it far inferior generalization performance. Because the outliers directly impact the margin, SVM and RLR are somewhat inferior to DWD, whose "weighted influence of all observations" allows better adaptation (here the difference is generally statistically significant, in the sense that 3 of the 5 pairs of confidence intervals don't overlap).

Figure 2(d) compares performance of these methods for the *nested spheres* data. This example is intended to study the relative performance of these methods for highly non-Gaussian distributions, as opposed to the relatively minor departure from normality that drove the above examples. Here the first $d/2$ dimensions are chosen so that Class -1 data are standard Gaussian, and Class $+1$ data are $\left[\frac{1+2.2\sqrt{2/d}}{1-2.2\sqrt{2/d}}\right]^{1/2}$ times standard Gaussian. This part of the data represent the perhaps canonical example of data that are very hard to separate by hyperplanes (a simplifying assumption of this paper). This time, the polynomial embedding method based on the ideas of

Aizerman, Braverman and Rozoner (1964), is considered, which is done by taking the remaining $d/2$ entries of each data vector to be the squares of the first $d/2$. This provides a path to very powerful discrimination, because linear combinations of the entries includes the sum of the squares of the first $d/2$ coordinates, which has excellent discriminatory power.

Because all of MD, RLR, SVM and DWD can find the sum of squares (i.e., realize that the discriminatory power lies in the second half of the data), it is not surprising that all give quite acceptable performance, although RLR lags somewhat for large d (Figure 2(d)). Despite the non-normality of this setting, MD is surprisingly the best of the methods for higher dimensions d (we don't know why, but believe it is related to this special geometric structure). Also unclear is why SVM is best only for dimension 10. Perhaps less surprising is that DWD is "in between" in the sense of being best for intermediate dimensions. The key to understanding these phenomena may lie in understanding how "data piling" works in polynomial embedded situations.

Note that in all the examples, most methods tend to come together at the right edge of each summary plot. This effect is explained by the HDLSS asymptotics of Hall, Marron and Neeman (2005) and Ahn, Marron, Muller and Chi (2007), where it is seen that under appropriate assumptions multivariate data tend (in the limit as $d \to \infty$, with n_+ and n_- fixed) to have a rigid underlying geometric structure, while all of the randomness appears in random rotations of that structure. Those asymptotics are also used for a mathematical statistical analysis of SVM and DWD in the former paper. It is also shown that in this HDLSS limit, all discrimination rules tend to give similar performance to the first order as observed here. However, a deeper look at the convergence to that limit revealed important differences between methods, which are consistent with the motivation for DWD given here.

We have also studied other examples. These are not shown to save space, and

because the lessons learned in the other examples are fairly similar. Figure 2(d) is a good summary: each method is best in some situations, and the special strength of DWD comes from its ability to frequently mimic the performance of the best ones.

4   Real Data Examples

We have studied two real data examples. The first is the microarray gene expression data from Perou et al. (1999). The data are vectors representing relative expression of $d = 456$ genes (chosen from a larger set as discussed in Perou et al. (1999)), from breast cancer patients. Because there are only $n = 136$ total cases available, this is a HDLSS setting. There are separate training and testing data sets. Here we consider 4 groups of binary classification problems, chosen for biological interest.

Group 1) Luminal cancer vs. other cancer types and normals. (85 training and 51 test cases).

Group 2) Luminal A vs. Luminal B&C. (50 training and 21 test cases).

Group 3) Normal vs. Erb & Basal cancer types. (38 training and 30 test cases).

Group 4) Erb vs. Basal cancer types. (25 training and 21 test cases).

MD has no tuning parameter, and the other three methods were tuned by leave-one-out cross-validation on the training data. Figure 3 shows the misclassification rates.

All four classification methods give overall reasonable performance. For groups 1 and 4, all methods give very similar good performance. Differences appear for the other groups, DWD and RLR being clearly superior for Group 2, but DWD is the worst of the four methods (although not by much) for Group 3. The overall lessons here are representative of our experience with other data analyses. Each method seems to have situations where it works well, and others where it is inferior. The promise of the DWD method comes from its very often being competitive with the
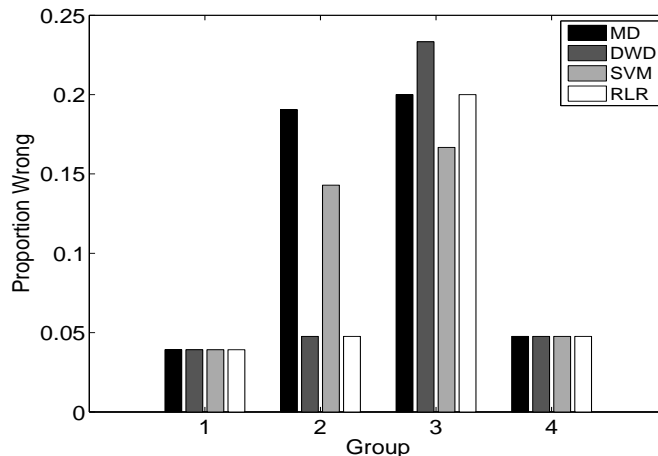
Figure 3: Graphical summary of classification error rates for gene expression data.

best of the others, and sometimes being better.

We also analyzed the Wisconsin diagnostic breast cancer data (Street, Wolberg and Mangasarian 1993). Details can be found in Section 4 of the second version of the paper, available at Ahn (2007), but the main lesson is consistent with above: each of these methods has the potential to be quite effective, and their relative differences are not large. Although DWD specifically targets HDLSS setting, it is good to see effective performance in other settings as well.

References

[1] Ahn, J. (2007), Distance Weighted Discrimination:

http://www.stat.uga.edu/∼jyahn/DWD/.

[2] Ahn, J. and Marron, J. S. (2005), "The direction of maximal data piling in high dimensional spaces," *Technical Report, University of North Carolina at Chapel Hill.*

[3] Ahn, J., Marron, J. S., Muller, K. M., and Chi Y. -Y. (2007), "The high dimension, low sample size geometric representation holds under mild conditions," to appear in *Biometrika*.

[4] Aizerman, M., Braverman, E., and Rozoner, L. I (1964), "Theoretical foundations of the potential function method in pattern recognition," *Automation and Remote Control*, 15, 821–837.

[5] Alizadeh, F. and Goldfarb, D. (2003), "Second-order cone programming," *Mathematical Programming*, 95, 3–51.

[6] 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*, 20, 105–114.

[7] le Cessie, S. and van Houwelingen, J. C. (1992), "Ridge estimators in logistic regression," *Applied Statistics*, 41, 191–201.

[8] Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, United Kingdom.

[9] Dandurova, Y., Yeganova, L., and Falk, J. E. (2001), "Robust set separation via exponentials," *Nonlinear Analysis*, 47, 1893–1904.

[10] Gunn, S. R. (1997), "Support Vector Machines for Classification and Regression," *Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton*.

[11] 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, 427–444 .

[12] Perou, C. M., Jeffrey, S. S., van de Rijn, M., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Rees, C. A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999), "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," *Proceedings of the National Academy of the Sciences, U.S.A.* 96, 9212–9217.

[13] Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993), "Nuclear feature extraction for breast tumor diagnosis," *IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology*, 1905, 861–870.

[14] Tütüncü, R. H., Toh, K. C., and Todd, M. J. (2003), "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming*, 95, 189–217.

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