# Distance Weighted Discrimination

J. S. Marron
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853
and
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
Email: marron@email.unc.edu

Michael Todd
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853
Email: miketodd@cs.cornell.edu

Jeongyoun Ahn
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
Email: jyahn@email.unc.edu

October 26, 2004

**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 (often much larger) than the sample size $n$ (the number of data vectors available). There is a strong need for HDLSS methods in the areas of genetic micro-array analysis (usually a very few cases, where many gene expression levels have been measured, see for example Perou et al. (1999)), chemometrics (typically a small population of high dimensional spectra, see for example Marron, Wendelberger and Kober (2004)) and medical image analysis (a small population of 3-d shapes represented by vectors of many parameters, see for example Yushkevich, Pizer, Joshi and Marron (2001) and Koch, Marron and Chen (2004)), and text classification (here there are typically many more cases, and also far more features, see for example Zhang and Oles (2001) and Peng and McCallum (2004)). Classical multivariate analysis is useless (i.e. fails completely to give a meaningful analysis) in HDLSS contexts, because the first step in the traditional approach is to *sphere the data*, by multiplying by the root inverse of the covariance matrix, which does not exist (because the covariance is not of full rank). Thus HDLSS settings are a large fertile ground for the re-invention of almost all types of statistical inference.

In this paper, the focus is on two class discrimination, with class labels $+1$ and $-1$. A clever and powerful discrimination method is the Support Vector Machine (SVM), proposed by Vapnik (1982, 1995). The SVM is introduced graphically in Figure 1 below. See Burges (1998) for an easily accessible introduction. It is useful to understand the SVM from a number of other additional viewpoints as well, see Cristianini and Shawe-Taylor (2000), Hastie, Tibshirani and Friedman (2001) and Schölkopf and Smola (2002). See Howse, Hush and Scovel (2002) for a recent overview of related mathematical results, including performance bounds. The first contribution of the present paper is a novel view of the performance of the SVM, in HDLSS settings, via projecting the data onto the normal vector of the separating hyperplane. This view reveals substantial *data piling*, which means that many of these projections (onto the normal direction vector) are the same, i. e. the projection coefficients are identical. For the 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 (as seen below in Figure 3). The extreme example of data piling appears in Figure 2.

The discussion around Figures 2 and 3 below suggests that data piling may adversely affect the *generalization performance* (how well new data from the same distributions can be discriminated) of the SVM in at least some HDLSS situations. The major contribution of this paper is a new discrimination method, called "Distance Weighted Discrimination" (DWD), which avoids the data piling problem, and is seen in the simulations in Section 3 to give the anticipated improved generalizability. Like the SVM, the computation of the DWD is based on computationally intensive optimization, but while the SVM uses well-known quadratic programming algorithms, the DWD uses recently developed interior-point methods for so-called Second-Order Cone Programming (SOCP) problems, see Alizadeh and Goldfarb (2003), discussed in detail in Section 2.2. The improvement available in HDLSS settings from the DWD comes from solving an optimization problem which yields improved data piling properties, as shown in Figure 4 below.

The two-class discrimination problem begins with two sets (classes) of $d$-dimensional training data vectors. A toy example, with $d = 2$ for easy viewing of the data vectors via a scatterplot, is given in Figure 1. The first class, called "Class $+1$," has $n_+ = 15$ data vectors shown as red plus signs, and the second class, called "Class $-1$," has $n_- = 15$ data vectors shown as blue circles. 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 $-1$." In this paper, it is assumed that the Class $+1$ vectors are independent and identically distributed random vectors from an unknown multivariate distribution (and similarly, but from a different distribution, for the Class $-1$ vectors).

For simplicity only *linear* discrimination methods are considered here. Note that "linear" is not meant in the common statistical sense of "linear function of the training data" (in fact most methods considered here are quite nonlinear as functions of the training data). Instead this means that the discrimination rule is a simple linear function of the new (testing) data vector to be classified. In particular, there is a direction vector $w$, and a threshold $\beta$, so that the new data vector $x$ is assigned to the Class $+1$ exactly when $x'w + \beta \geq 0$. This corresponds to separation of the $d$-dimensional data space into two regions by a hyperplane, with normal vector $w$, whose position is determined by $\beta$. In Figure 1, one such normal vector $w$ is shown as the thick purple line, and the corresponding separating hyperplane (in this case having dimension $d = 2 - 1 = 1$) is shown as the thick green dashed line. Extensions to the *nonlinear* case are discussed at various points below.

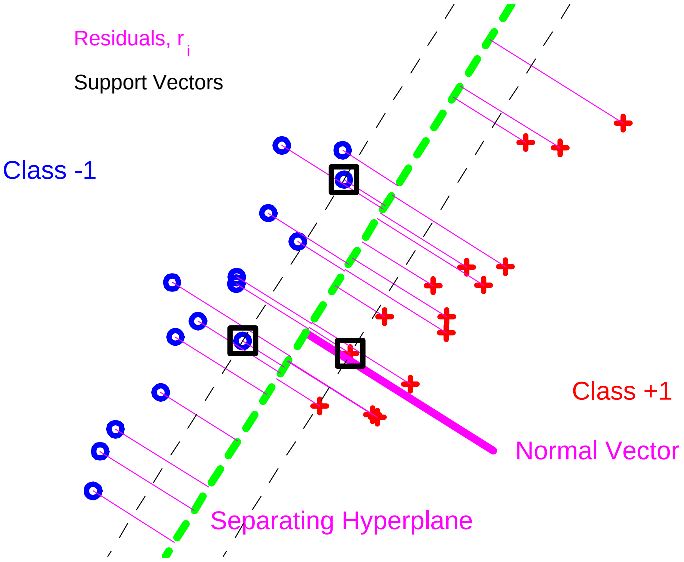

FIGURE 1: *Toy Example illustrating the Support Vector Machine. Class +1 data shown as red plus signs, and Class $-1$ data shown as blue circles. The separating hyperplane is shown as the thick dashed line, with the corresponding normal vector shown as the thick solid line. The residuals, $r_i$, are the thin lines, and the support vectors are highlighted with black boxes.*

The separating hyperplane shown in Figure 1 does an intuitively appealing job of separating the two data types. This is the SVM hyperplane, and the remaining graphics illustrate how it was constructed. The key idea behind the SVM is to find $w$ and $\beta$ to keep the data in the same class all on the same side of, and also as far as possible from, the separating hyperplane (in an appropriate sense). This is quantified using a maximin optimization formulation, focussing on only the data points that are closest to the separating hyperplane, called *support vectors*, highlighted in Figure 1 with black boxes. The hyperplanes parallel to the separating hyperplane that intersect the support vectors are shown as thin black dashed lines. The distance between these hyperplanes is called the *margin*. The SVM finds the separating hyperplane that maximizes the margin, with the solution for these data being shown in Figure 1. An alternative view is that we find two closest points, one in the convex hull of the Class +1 points and one in the convex hull of the Class $-1$ points. The SVM separating hyperplane will then be the perpendicular bisector of the line segment joining two such points, i.e., this line segment is parallel to the normal vector shown in Figure 1. Note that the convex combinations defining these closest points only involve the support vectors of each class.

Figure 1 also shows how the SVM can be subject to data piling, in the direction of the normal vector. As defined above, given a direction vector, *data piling* occurs when many data points have identical projections in that direction, i.e. the data pile up on top of each other. A simple extreme example of data piling direction vector is one that is orthogonal to the subspace generated by the data (such directions are always available in HDLSS settings), because the projections are identically 0. Another complete data piling direction is any normal vector to the hyperplane generated by the data, where data all project on a single value that is typically non-zero. Of more relevance for our purposes is a somewhat different type of data piling, that is relevant to discrimination, where (for some given direction vectors) at least some of the data points for each class pile at two different common points, one for each class. Figure 1 shows that the SVM has this potential, in the case where there are many support vectors (i.e., points lying on the boundary planes shown as dashed black lines in Figure 1). In particular, when the data are projected onto the normal vector, because the boundary planes are orthogonal, the support vectors will all be projected to one of two common points whose distance to the Separating Plane is the margin. While the number of support vectors is small in the 2 dimensional example shown in Figure 1, it can be quite large in HDLSS settings, as shown in Figure 3 below.

The toy example in Figure 1 is different from the HDLSS focus of this paper because the sample sizes $n_+$ and $n_-$ are larger than the dimension $d = 2$. Some perhaps surprising effects occur in HDLSS contexts. This point is illustrated in Figure 2. The data in Figure 2 have dimension $d = 39$, with $n_+ = 20$ data vectors from Class $+1$ represented as red plus signs, and $n_- = 20$ data vectors from Class $-1$ represented as blue circles. The data were drawn from 2 distributions that are nearly standard normal (i.e., Gaussian with zero mean vector and identity covariance), except that the mean in the first dimension only is shifted to +2.2 (-2.2 resp.) for Class $+1$ ($-1$ resp.). The data are not simple to visualize because of the high dimension, but some important lower dimensional projections are shown in the various panels of Figure 2.

The thick, dashed purple line in Figure 2 shows the first dimension. Because the true difference in the Gaussian means lies *only* in this direction, this is the normal vector of the theoretical Bayes risk optimal separating hyperplane. Discrimination methods whose normal vector lies close to this direction should have good generalization properties, i.e., new data will be discriminated as well as possible. The thick purple line is 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 the property that both classes pile completely onto just two points, one for each class (i.e., all projections on that data vector are identical to one value for each class). Out of all of these, there is a unique direction, which maximizes the separation of the two piling points, called the *maximal data piling* direction. This direction is computed as $w = \widehat{\Sigma}^{-1}(\overline{x}^+ - \overline{x}^-)$, where $\overline{x}^+$ ($\overline{x}^-$ resp.) is the mean vector of Class $+1$ ($-1$ resp.), and $\widehat{\Sigma}$ represents the covariance matrix of the full data set, with the superscript $-1$ indicating the generalized inverse. The generalized inverse is needed in HDLSS situations, because the covariance matrix is not of full rank. The direction $w$ is nearly that of Fisher Linear Discrimination, except that it uses the full data covariance matrix, instead of the within class version. As noted by a referee, the true Fisher Linear Discrimination direction (based on the pooled within class covariance matrix) does not give data piling, see Figure 15.3 in Schölkopf and Smola (2002). In Ahn and Marron (2004) it is seen that the two methods are identical in non-HDLSS situations. However Bickel and Levina (2004) have demonstrated very poor HDLSS properties of the true FLD, which is consistent with the corresponding version of Figure 2 showing an even worse angle with the optimal vector (not shown here to save space). The first dimension, together with the vector $w$, determine a two-dimensional subspace, and the top panel of Figure 2 shows the projection of the data onto that two-dimensional plane. Another way to think of the top panel is that the full $d = 39$-dimensional space is rotated around the axis determined by the first dimension, until the two dimensional plane contains the vector $w$. Note that the data within each class appear to be collinear.

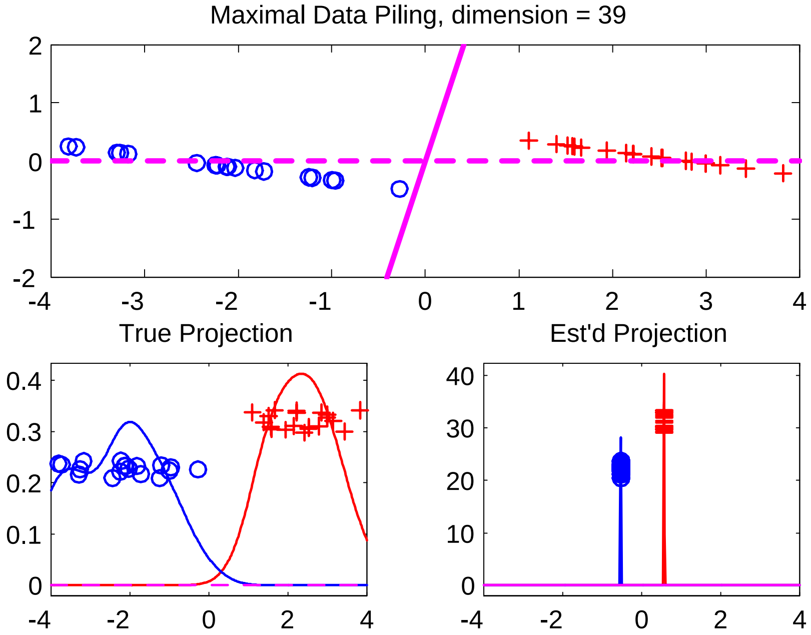

FIGURE 2: *Toy Example, illustrating potential for "data piling" problem in HDLSS settings. Dashed purple line is the Bayes optimal direction, solid is the Maximal Data Piling direction. Top panel is two-dimensional projection ($x$-axis is in the optimal direction, plane is determined by the MDP vector), bottom panels are one-dimensional projections ($x$-axis shows projections, $y$-axis is density).*

Other useful views of the data include one-dimensional projections, shown in the bottom panels. The bottom left is the projection onto the true optimal discrimination direction, shown as the dashed line in the top panel. The bottom right shows the projection onto the direction $w$, which is the solid line in the top panel. In both cases, 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 included in the bottom panel are kernel density estimates, which give another indication of the structure of these univariate populations. As expected, the left panel reveals two Gaussian populations, with respective means ±2.2. The bottom right panel shows that indeed the data essentially line up in a direction orthogonal to the solid purple line, resulting in *data piling*.

Data piling is usually not a useful property for a discrimination rule, because the corresponding direction vector 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, which are typically expected to bear no relation to these, and thus will result in a completely different maximal data piling direction. In this sense, the maximal data piling direction will typically not be generalizable, because it is driven by noise artifacts of just the data at hand. Another way of understanding this comes from study of the solid direction vector $w$ in the top panel. The poor generalization properties are shown by the fact that it is not far from orthogonal to the optimal direction vector shown as the dashed line. Projection of a new data vector onto $w$ cannot be expected to provide effective discrimination, because of the arbitrariness of the direction $w$.

It is of interest to view how Figure 2 changes as the dimension changes. This can be done by viewing the movie in the file DWD1figMDP.avi available in the web directory Marron (2004). This shows the same view for selected dimensions $d = 1, ..., 1000$. For small $d$, the solid line is not far from the dashed line, but data piling begins as $d$ approaches $n_+ + n_- - 1 = 39$. Past that threshold the points pile up perfectly, and then the two piles slowly separate, since for higher $d$, there are more "degrees of freedom of data piling." The worst performance is for $d \approx n_+ + n_-$, and performance is seen to actually improve as $d$ grows beyond that level. This is consistent with the HDLSS asymptotics of Hall, Marron, and Neeman (2004), where it is seen that under standard 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 that paper. It is also shown that in this HDLSS limit, all discrimination rules tend to give similar performance to the first order (this is observed in our simulations in Section 3), so that the maximal data piling discrimination rule gives reasonable performance in this limit.

The data piling properties of the SVM are studied in Figure 3. Both the data, and also the graphical representation, are the same as in Figure 2. The only difference is that now the direction $w$ is determined by the SVM. The top panel shows that the direction vector $w$ (the solid line) is already much closer to the optimal direction (the dashed line) than for Figure 1. This reflects the reasonable generalizability properties of the SVM in HDLSS settings. The SVM is far superior to the maximal data piling modification shown in Figure 2, because the normal vector, shown as the thick purple line, is much closer to the

Bayes optimal direction (recall these were nearly orthogonal in Figure 2), shown as the dashed purple line. However the bottom right panel suggests that there is room for improvement. In particular, there is a clear piling up of data at the margin. As in Figure 2 above, this shows that the SVM is affected by spurious properties of this particular realization of the training 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). Again, a richer visualization of this phenomenon can be seen in the movie version in the file DWD1figSVM.avi in Marron (2004). The improved generalizability of the SVM is seen over a wide range of dimensions.

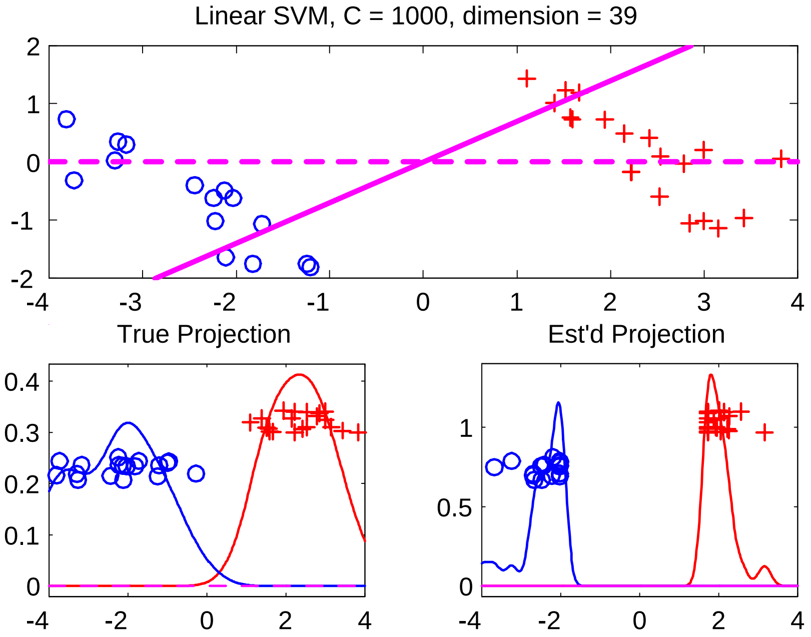

Figure 3: *Same toy example illustrating partial "data piling" present in HDLSS situations, for discrimination using the Support Vector Machine. Format is same as Figure 2.*

Room for improvement of the generalizability of the SVM, in HDLSS situations, comes from allowing more of the data points (beyond just those on the margin) to have a direct impact on the direction vector $w$. In Section 2.2 we propose the new Distance Weighted Discrimination method. Like the SVM, this is the solution of an optimization problem. However, the new optimization replaces the maximin "margin based" criterion of the SVM, by a different function of the distances, $r_i$, from the data to the separating hyperplane, shown as thin purple lines in Figure 1. A simple way of allowing these distances to influence the direction $w$ 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. Additional insight comes from an alternative (dual) view. The normal to the separating hyperplane is again the difference between a convex combination of the Class $+1$ points and a convex combination of the Class $-1$ points, but now the combinations are chosen to minimize the distance between the points divided by the square of the sum of the square roots of the weights used in the convex combination. In this way, all points receive a positive weight.

The difference between the two solutions can be seen in a very small example. Suppose there is just one Class $+1$ point, $(3; 0)$, and four Class $-1$ points, $(-3; 3)$, $(-3; 1)$, $(-3; -1)$, and $(-3; -3)$. (We use Matlab-style notation, so that $(a; b)$ denotes the vector with vectors or scalars $a$ and $b$ concatenated into a single column vector, etc.) The SVM maximizes the margin and gives $(1, 0)x = 0$ as the separating hyperplane. The DWD has four points on the left "pushing" on the hyperplane and only one on the right (we are using the mechanical analogy explained more in Section 2), and the result is that the separating hyperplane is translated to $(1, 0)x - 1 = 0$. Note that the "class boundary" is at the mid-point value of 0 for the SVM, while it is at the more appropriately weighted value of 1 for the DWD. The SVM class boundary would be more appealing if the unequal sample numbers are properly taken into account, but adding three Class $+1$ points around $(100; 0)$ equalizes the class sizes and leaves the result almost unchanged (because the new points are so far from the hyperplane).

Let us now return to the example shown in Figures 2 and 3. The DWD version of the normal vector is shown as the solid line in Figure 4. Note that this is much closer to the Bayes optimal direction (shown as the dashed line), than for either the maximal data piling direction shown in Figure 2, or the SVM shown in Figure 3. The lower right hand plot shows no "data piling," which is the result of each data point playing a role in finding this direction in the data.

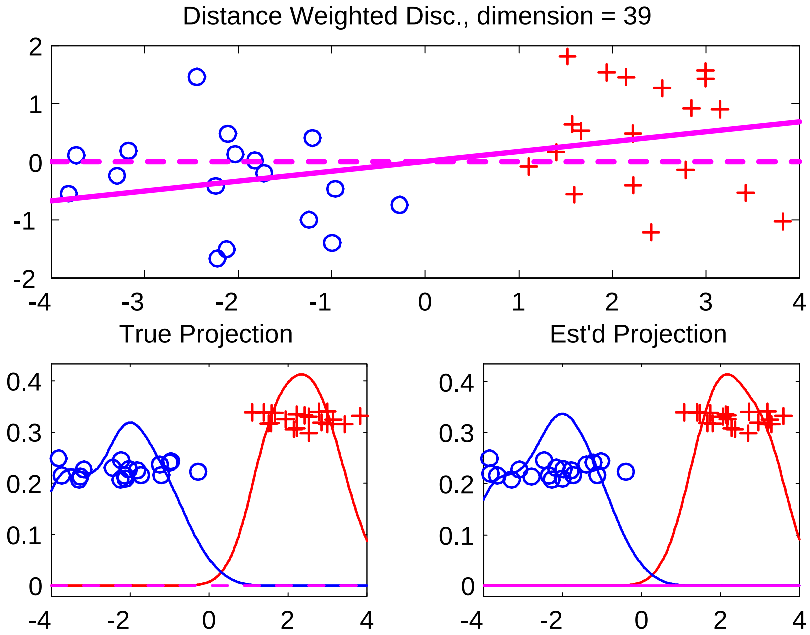

FIGURE 4: *Same Toy Example, illustrating no "data piling" for Distance Weighted Discrimination. Format is same as Figure 2.*

Once again, the corresponding view in a wide array of dimensions is available in the movie version in file DWD1figDWD.avi in the web directory Marron (2004). This shows that the DWD gives excellent performance in this example over a wide array of dimensions.

An additional advantage of the approximately Gaussian sub-population shapes shown in the lower right of Figure 4, compared to the data piled, approximately triangular, shapes shown in the lower right of Figure 3, is that DWD then provides immediate and natural application to micro-array bias adjustment (where the two sub-populations are moved along this vector until they overlap to remove data biases), as done in Benito et al. (2004).

Note that all of these examples show "separable data," where there exists a hyperplane which completely separates the data. This is typical in HDLSS settings, but is not true for general data sets, see e.g. Cover (1965). Both SVM and DWD approach this issue (and also potential gains in generalizability) via an extended optimization problem, which incorporates penalties for "violation" (i.e., a data point being on the wrong side of the separating hyperplane). Some price is incurred by this approach, because it requires selection of a tuning parameter.

Precise formulations of the optimization methods that drive SVM and DWD are given in Section 2. This includes discussion of computational complexity issues in Section 2.3, and tuning parameter choice in Section 2.4. Simulation results, showing the desirable generalization properties of DWD, and contrasting it with some important competitors, are given in Section 3. 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 that of the SVM when it is superior, and also close to that of the simple mean difference method in settings where it is best. Similar overall performance of DWD is also shown on micro-array and breast cancer data sets in Section 4. Some open problems and future directions are discussed in Section 5, and Section 6 is an appendix giving further details on the optimization problems.

A side issue is that from a purely algorithmic viewpoint, one might wonder: why do HDLSS settings require unusual treatment? For example, even when $d \gg n$, the data still lie in an $n$ dimensional subspace, so why not simply work in that subspace? The first answer is that new data are expected to fall outside of this subspace, so that restriction to this subspace will not allow proper consideration of generalizability, which is an intrinsically $d$-dimensional notion. The second answer, is that in the very important HDLSS context of micro-arrays for gene expression, there is typically interest in some subsets of specific genes. This focus on subsets is much harder to do when *only* a few linear combinations (i.e. any basis of the subspace generated by the data) of the genes are considered.

## 2 Formulation of Optimization Problems

This section gives details of the optimization problems underlying the original Support Vector Machine, and the Distance Weighted Discrimination ideas proposed here.

Let us first set the notation to be used. The training data consists of $n$ $d$-vectors $x_i$ together with corresponding class indicators $y_i \in \{+1, -1\}$. We let $X$ denote the $d \times n$ matrix whose columns are the $x_i$'s, and $y$ the $n$-vector of the $y_i$'s. The two classes of Section 1 are both contained in $X$, and are distinguished using $y$. Thus, the quantities $n_+$ and $n_-$ from Section 1 can be written as: $n_+ = \sum_{i=1}^n 1_{\{y_i=+1\}}$ and $n_+ = \sum_{i=1}^n 1_{\{y_i=-1\}}$, and we have $n = n_+ + n_-$. It is convenient to use $Y$ for the $n \times n$ diagonal matrix with the components of $y$ on its diagonal. Then, if we choose $w \in \Re^d$ as the normal vector (the thick solid purple line in Figure 1) for our hyperplane (the thick dashed green line in Figure 1) and $\beta \in \Re$ to determine its position, the residual of the $i$th data point (shown as a thin solid purple line in Figure 1) is

$$\bar{r}_i = y_i(x_i'w + \beta),$$

or in matrix-vector notation

$$\bar{r} = Y(X'w + \beta e) = YX'w + \beta y,$$

where $e \in \Re^n$ denotes the vector of ones. We would like to choose $w$ and $\beta$ 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 $w$ and $\beta$, so $w$ 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 $\xi \in \Re^n_+$ of errors, to be suitably penalized, and define the perturbed residuals to be

$$r = YX'w + \beta y + \xi. \tag{1}$$

When the data vector $x_i$ lies on the proper side of the separating hyperplane and the penalization is not too small, $\xi_i = 0$, and thus $\bar{r}_i = r_i$. Hence the notation in Figure 1 is consistent (i.e., there is no need to replace the label $r_i$ by $\bar{r}_i$).

The SVM chooses $w$ and $\beta$ to maximize the minimum $r_i$ in some sense (details are given in Section 2.1), while our Distance Weighted Discrimination approach instead minimizes the sum of reciprocals of the $r_i$'s augmented by a penalty term (as described in Section 2.2). Both methods involve a tuning parameter that controls the penalization of $\xi$, whose choice is discussed in Section 2.4.

While the discussion here is mostly on "linear discrimination methods" (i.e., those that attempt to separate the classes with a hyperplane), it is important to note that this actually entails a much larger class of discriminators, through "polynomial embedding" and "kernel embedding" ideas. 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. An important point is that most methods that are sensible for the simple linear problem described here are also viable in polynomial or kernel embedded contexts as well, including not only the SVM and DWD, but also perhaps more naive methods such as Fisher Linear Discrimination.

## 2.1 Support Vector Machine Optimization

For general references, see Cristianini and Shawe-Taylor (2000), Hastie, Tibshirani and Friedman (2001) and Schölkopf and Smola (2002).

Suppose first that the data are strictly linearly separable and that no perturbations are used. Then maximizing the minimum residual can be achieved by introducing a new variable $\delta$ and maximizing $\delta$ subject to $\bar{r} \geq \delta e$. As noted above, $\bar{r}$ scales with $w$ and $\beta$, so instead of restricting the norm of $w$ to 1 and maximizing $\delta$ (which can be made positive), we can instead restrict $\delta$ to 1 and minimize the norm of $w$, or equivalently half the norm of $w$ squared, to get a convex quadratic function. Now we allow nonnegative perturbations $\xi$, and add a penalty on the 1-norm of $\xi$ to the objective function of this minimization problem. The result is the (soft-margin) SVM problem

$$(P_{SVM}) \qquad \min_{w,\beta,\xi} \quad (1/2)w'w + Ce'\xi, \quad YX'w + \beta y + \xi \geq e, \; \xi \geq 0.$$

where $C = C_{SVM} > 0$ is a penalty parameter.

This convex quadratic programming problem has a dual, which turns out to be

$$(D_{SVM}) \qquad \max_{\alpha} \quad -(1/2)\alpha' Y X' X Y \alpha + e'\alpha, \quad y'\alpha = 0,\, 0 \leq \alpha \leq Ce,$$

and both problems have optimal solutions. Under suitable conditions (which hold here), the two problems have equal optimal values, and solving one allows you to find the optimal solution to the other.

Section 6.1 in the appendix 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, and how the SVM can be extended to the nonlinear case using a kernel function (Burges (1998), Section 4, or Cristianini and Shawe-Taylor (2000), Chapter 3).

From the optimality conditions we can see that, if all $x_i$'s are scaled by a factor $\gamma$, then the optimal $w$ is scaled by $\gamma^{-1}$ and the optimal $\alpha$ by $\gamma^{-2}$. It follows that the penalty parameter $C$ should also be scaled by $\gamma^{-2}$. Similarly, if each training point is replicated $p$ times, then $w$ remains the same while $\alpha$ is scaled by $p^{-1}$. Hence a reasonable value for $C$ is some large constant divided by $n$ times a typical distance between $x_i$'s squared. The choice of $C$ is discussed further in Section 2.4.

## 2.2 Distance Weighted Discrimination Optimization

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 $\xi$, be minimized: thus we have

$$\min_{r,w,\beta,\xi} \quad \sum_i (1/r_i) + Ce'\xi, \quad r = YX'w + \beta y + \xi,\, (1/2)w'w \leq 1/2,\, r \geq 0,\, \xi \geq 0,$$

where again $C = C_{DWD} > 0$ is a penalty parameter. (We have relaxed the condition that the norm of $w$ 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, the optimal solution will have the norm equal to 1.)

In Section 6.2 in the appendix, 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

$$S_{m+1} := \{(\zeta; u) \in \Re^{m+1} : \zeta \geq \|u\|\}.$$

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. After some manipulations, we arrive at

$$\begin{array}{llcccccccccccl}
\min_{\psi,w,\beta,\xi,\rho,\sigma,\tau} & & & & & & Ce'\xi & + & e'\rho & + & e'\sigma & & & \\
& & YX'w & + & \beta y & + & \xi & - & \rho & + & \sigma & & = & 0, \\
& \psi & & & & & & & & & & & = & 1, \\
(P_{DWD}) & & & & & & & & & & & \tau & = & e,
\end{array}$$

$$(\psi; w) \in S_{d+1}, \quad \xi \geq 0, \quad (\rho_i; \sigma_i; \tau_i) \in S_3,\ i = 1, 2, \ldots, n.$$

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

$$(D_{DWD}) \qquad \max_{\alpha} \quad -\|XY\alpha\| + 2e'\sqrt{\alpha}, \quad y'\alpha = 0,\ 0 \leq \alpha \leq Ce.$$

(Here $\sqrt{\alpha}$ denotes the vector whose components are the square roots of those of $\alpha$.) Compare with $(D_{SVM})$ above, which is identical except for having objective function $-(1/2)\|XY\alpha\|^2 + e'\alpha$. Again, both problems have optimal solutions.

Section 6.2 in the appendix 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 the DWD can again be extended to the nonlinear case using a kernel function.

Problem $(P_{DWD})$ has $2n + 1$ equations and $3n + d + 2$ variables, and so solving it can be expensive in the large-scale HDLSS case. This is discussed further in the next subsection. The appendix also shows how a preprocessing step can be performed to reduce $d$ to $n$.

From the optimality conditions, or from the interpretation of $C$ in the appendix, we can see that, if all $x_i$'s are scaled by a factor $\gamma$, the penalty parameter $C$ should be scaled by $\gamma^{-2}$, 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 $x_i$'s squared.

## 2.3 Computational Complexity

Problems $(P_{SVM})$ and $(D_{SVM})$ are standard convex quadratic programming problems, for which many efficient algorithms exist. Primal-dual interior-point methods can be used, for example, which guarantee to obtain $\epsilon$-optimal solutions to both problems starting with suitable initial points in $O(\sqrt{n}\ln(1/\epsilon))$ iterations, where each iteration requires the solution of a square linear system of dimension $n$; in practice, these methods obtain a solution to within machine precision within $10 - 50$ iterations; see, for example, Wright (1997). We can view each iteration as a Newton-like step to solve barrier problems associated with the primal and dual problems, or alternatively to solve perturbed versions of the optimality conditions for these problems. However, there are also other methods, such as active-set methods and iterative gradient-based methods, which lack such guarantees but typically solve such problems very efficiently, particularly when applied to ($D_{SVM}$) when the number of positive $\alpha_i$'s may be small. See, e.g., Cristianini and Shawe-Taylor (2000), Chapter 7.

For the SOCP problems ($P_{DWD}$) and ($P_{DWD}$), there are again 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. Once again, a small number of iterations is required (theoretically $O(\sqrt{n}\ln(1/\epsilon))$, 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. Again, 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 Section 6.2 in the appendix, 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. Note that, because each $\alpha_i$ is positive in the optimal solution, techniques like chunking and decomposition (see Chapter 7 in Cristianini and Shawe-Taylor (2000)) are not useful.

## 2.4 Choice of tuning parameter

A simple recommendation for the choice of the tuning parameter $C$ is made here. It is important to note that this recommendation is intended for use in HDLSS settings, and even in that case there may be benefits to careful tuning of $C$. For non-HDLSS situations, or if careful tuning is desired, we recommend the cross-validation type ideas of Wahba, Lin, Lee and Zhang (2001) and Lin, Wahba, Zhang, and Lee (2002). We believe this is important because in non-HDLSS situations, it may be too much to hope that the data are separable, so one will be compelled to deal with violators (points on the wrong side of the separating hyperplane).

For both SVM and DWD, the above simple considerations suggest that $C$ should scale with the inverse square of the distance between training points, and in the SVM case, inversely with the number of training points. This will result in a choice that is essentially "scale invariant," i.e., if the data are all multiplied by a constant, or replicated a fixed number of times, the discrimination rule will stay the same.

As a notion of *typical distance*, we suggest the median of the pairwise Euclidean distances between classes,

$$d_t = \text{median}\left\{\|x_i - x_{i'}\| : y_i = +1, y_{i'} = -1\right\}.$$

Other notions of typical distance are possible as well.

Then we recommend using "a large constant" divided by the typical distance squared, possibly divided by the number of data points in the SVM case. In most examples in this paper, we use $C = 100/d_t^2$, for DWD and we use Gunn's recommendation of $C = 1000$ for SVM. We view such simple use of defaults as important, because this is how most users will actually implement these methods. However, more careful tuning is also an important issue, so we have employed cross-validated tuning in the real data example of Section 4.1, and have studied a range of tuning parameters in Section 4.2. More careful choice of $C$ for DWD, in HDLSS situations, will be explored in an upcoming paper.

# 3 Simulations

In this section, simulation methods are used to compare the performance of DWD with the SVM. Also of interest is to compare both of these methods with the very simple "Mean Difference" (MD) method, defined in Section 3.1, and with the Regularized Logistic Regression method of le Cessie and van Houwelingen (1992), defined in Section 3.2.

These methods have been chosen because they are all do discrimination by finding a single separating hyperplane. Beyond the scope of our study is comparison with other methods, including those based on Nearest Neighbors and Neural Nets, and other approaches such as CART and MARS, see Duda, Hart and Stork (2000) for summarization of these.

## 3.1 The Mean Difference Discrimination Rule

The MD, also called the nearest centroid method (see for example Chapter 1 of Schölkopf and Smola (2002)) is a simple precursor to the shrunken nearest centroid method of Tibshirani et al (2002). It is based on the class sample mean vectors:

$$\begin{aligned}\overline{x}^{+} &= \frac{1}{n_{+}}\sum_{i=1}^{n} x_i 1_{\{y_i=+1\}},\\ \overline{x}^{-} &= \frac{1}{n_{-}}\sum_{i=1}^{n} x_i 1_{\{y_i=-1\}},\end{aligned}$$

and a new data vector is assigned to Class $+1$ ($-1$ resp.), when it is closer to $\overline{x}^{+}$ ($\overline{x}^{-}$ resp.). This discrimination method can also be viewed as attempting to find a separating hyperplane (as done by the SVM and DWD) between the two classes. This is the hyperplane with normal vector $\overline{x}^{+} - \overline{x}^{-}$, which bisects the line segment between the class means. Note that this compares nicely with the interpretations of the dual problems ($D_{SVM}$) and ($D_{DWD}$), where again the normal vector is the difference between two convex combinations of the Class $+1$ and Class $-1$ points. The MD is the maximum likelihood estimate of the (theoretical) Bayes Risk optimal rule for discrimination if the two class distributions are spherical Gaussian distributions (e.g. both have identity covariance matrices), and in a very limited class of other situations.

Fisher Linear Discrimination can be motivated by adjusting this idea to the case where the class covariances are the same, but of more complicated type. In classical multivariate settings (i.e., $n \gg d$), FLD is always preferable to

MD, because even when MD is optimal, FLD will be quite close, and there are situations (e.g. when the covariance structure is far from spherical) where the FLD is greatly improved. However, this picture changes completely in HDLSS settings. The reason is that FLD requires an estimate of the covariance matrix, based on a completely inadequate amount of data. This is the root of the "data piling" problem illustrated in Figure 2. In HDLSS situations the stability of MD gives it far better (even though it may be far from optimal) generalization properties than FLD. Bickel and Levina (2004) have pointed out that an important method that lies between FLD and MD (by taking scaling into account along individual coordinate axes) is commonly called "the naive Bayes method" in the machine learning literature.

MD is taken as the "classical statistical representative" in this simulation study. Its performance for the toy example considered in Section 1, can be seen in the movie DWD1figMD.avi, which is available in the web directory Marron (2004).

### 3.2 Regularized Logistic Regression

Classical logistic regression (for an overview, see Hastie, Tibshirani and Friedman (2001)) is a popular linear classification method and can be improved by adding a penalty term, controlled by a regularization parameter (le Cessie and van Houwelingen (1992)). In HDLSS cases, this regularization must be done to avoid ill-conditioning problems and to ensure better generalizability. It is called Regularized Logistic Regression (RLR) or penalized Logistic Regression. See Schimek (2003) for a recent application to gene expression analysis.

To simplify the formulation, it is convenient to change the possible values of $y_i$ from $\{-1, 1\}$ as done elsewhere in this paper, to $\{0, 1\}$. If we define $p(x)$ to be the probability of $y = 1$ given $x$, then the Bernoulli likelihood of $\{y_1, ..., y_n\}$ given $p(x_1), ..., p(x_n)$ is

$$L = \prod_{i=1}^{n} \left[ p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \right].$$

Using the logit link $g(x) = \log \frac{p(x)}{1-p(x)}$, the negative log-likelihood is

$$l := -\log L = \sum_{i=1}^{n} \left[ -y_i g(x_i) + \log(1 + e^{g(x_i)}) \right].$$

Linear RLR finds the separating hyperplane $g(x) = x'w + \beta$ which minimizes

$$\sum_{i=1}^{n} \left[ -y_i g(x_i) + \log(1 + e^{g(x_i)}) \right] + C \sum_{j=1}^{d} w_j^2,$$

where $w_j$ is the $j^{\text{th}}$ element of the vector $w$.

The first term is the negative log-likelihood and the second term is an $L^2$ penalty, which works in a fashion similar to ridge regression, where $C$ is the regularization parameter.

Some experimentation with the choice of $C$ suggested it did not have a large impact for the examples that we considered. In our simulation study, we used the SVM choice of $C = 1000$. We checked that the results of our simulation study (shown below) did not change over several values of $C$ ranging from 0.01 to $10,000$. In Section 4.1 careful choice of $C$ via cross-validation is considered, and explicit consideration of a range of choices is done in Section 4.2.

The RLR optimization problem can be solved by a Newton-Raphson iterative algorithm. In this paper, we used the "lrm" function in the S library "Design." Details about the program can be found at:
http://lib.stat.cmu.edu/S/Harrell/help/Design/html/00Index.html

## 3.3 Simulation Results

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 a wide array of HDLSS settings (from *not HDLSS* to *extremely HDLSS*). Each experiment was replicated 100 times. The graphics summarize the mean (over the 100 replications) of the proportion (out of the 200 members of each test data set) of incorrect classifications. To give an impression of the Monte Carlo variation, simple 95% confidence intervals for the mean value are also included as error bars.

The first distribution, studied in Figure 5, is essentially that of the examples shown in Figures 2-4. Both class distributions have unit covariance matrix, and the means are 0, except in the first coordinate direction, where the means are $+2.2$ ($-2.2$ resp.) for Class $+1$ ($-1$ resp.). Thus the (theoretical) Bayes Rule for this discrimination problem is to separate the classes with the hyperplane normal to the first coordinate axis. If it is known that one should look in the direction of the first coordinate axis, then the two classes are easy to separate, as shown in the bottom left panels of Figures 2-4. However, in high dimensions, it can be quite challenging to find that direction.

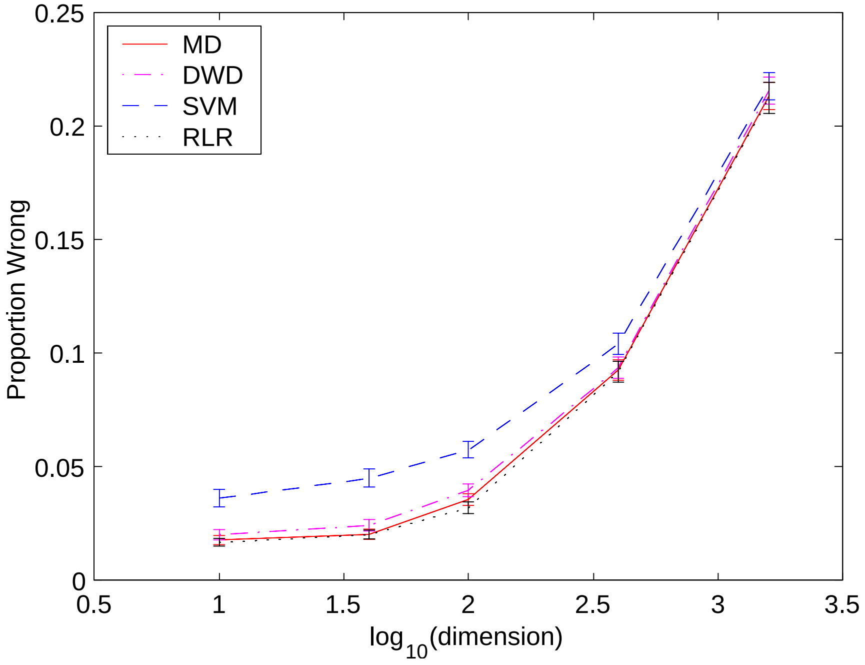

FIGURE 5: *Summary of simulation results for spherical Gaussian distributions. As expected, MD is the best, but not significantly better than DWD.*

The red curve in Figure 5 shows the generalizability performance of MD for this example. The classification error goes from about 2% for $d = 10$, to about 22% for $d = 1600$. For this example, the MD direction is the maximum likelihood estimate (based on the difference of the *sample* means) of the Bayes Risk optimal direction (based on the difference of the underlying *population* means), so the other methods are expected to have a worse error rate. RLR gives very similar performance, indeed being slightly better at $n = 100$, although the confidence intervals suggest the difference is not statistically significant. Note that the SVM, represented by the blue curve, has substantially worse error (the confidence intervals are generally far from overlapping), due to the data piling effect illustrated in Figure 3. However the purple curve, representing DWD, is much closer to optimal (the confidence intervals overlap). This demonstrates the gains that are available from DWD, relative to SVM, by explicitly using all of the data in choosing the separating hyperplane in HDLSS situations.

While the MD is the maximum likelihood estimate of the Bayes Risk optimal for spherical Gaussian distributions, it can be far from optimal in other cases. An example of this type, called the *outlier mixture distribution*, is a mixture distribution where 80% of the data are from the distribution studied in Figure 5, and the remaining 20% are Gaussian with mean +100 (−100 resp.) in the first coordinate, +500 (−500 resp.) in the second coordinate, and 0 in the other coordinates. Excellent discrimination for this distribution is again provided by the hyperplane whose normal vector is the first coordinate axis direction, because that separates the first 80% of the data well, and the remaining 20% are far away from the hyperplane (and on the correct side). Since the new 20% of the data will never be support vectors, SVM is expected to be similar to that in Figure 5. However, the new 20% of the data will create grave difficulties for the MD, because outlying observations have a strong effect on the sample mean, which will skew the normal vector towards the outliers, resulting in a poorly performing hyperplane. This effect is shown in Figure 6.

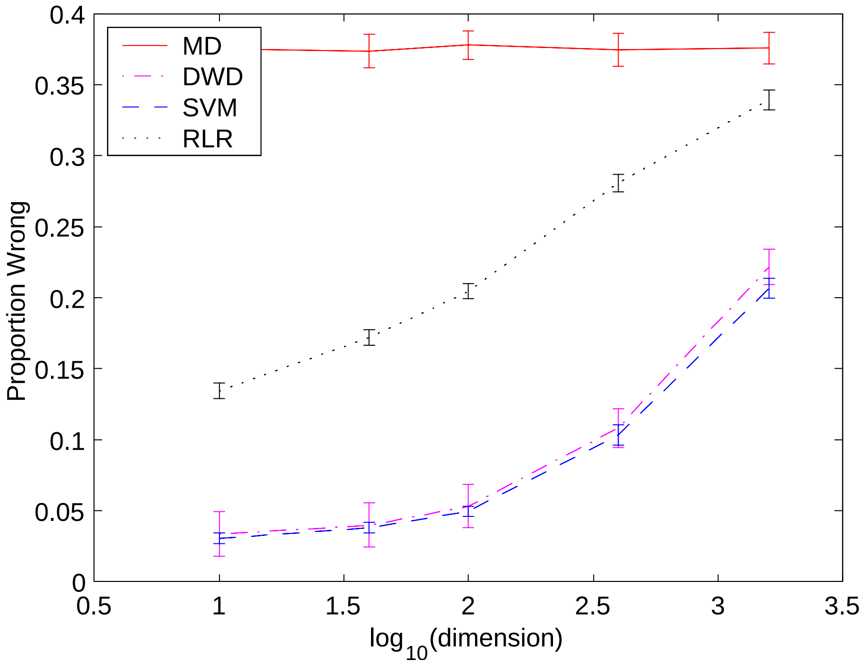

FIGURE 6: *Simulation comparison, for the outlier mixture distribution. SVM is the best method, but not significantly better than DWD.*

Note that in Figure 6, the SVM is best (as expected), because the outlying data are never near the margin. The MD has very poor error rate (recall that 50% error is expected from the classification rule which ignores the data, and instead uses a coin toss!), because the sample means are dramatically impacted by the 20% outliers in the data. RLR is similarly strongly affected by the outliers, while it is less sensitive than MD, it is clearly not as robust as SVM or DWD. DWD nearly shares the good properties of the SVM because the outliers receive a very small weight. While the DWD error rate is consistently above that for the SVM, lack of statistical significance of the difference is suggested by the overlapping error bars.

Figure 7 shows an example where the 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 distribution are from the shifted spherical

Gaussian as in Figure 5, and the remaining 20% are chosen so that the first coordinate is replaced by +0.1 (-0.1 resp.), and just one randomly chosen coordinate is replaced by +100 (-100, resp.), for an observation from Class +1 ($-1$, resp.). That is, a few pairs of observations are chosen to violate the ideal margin, in ways that push directly on the support vectors. Once again outliers are introduced, but this time, instead of being well away from the natural margin (as in the data that underlie the summary shown in Figure 6), they appear in ways that directly impact it.

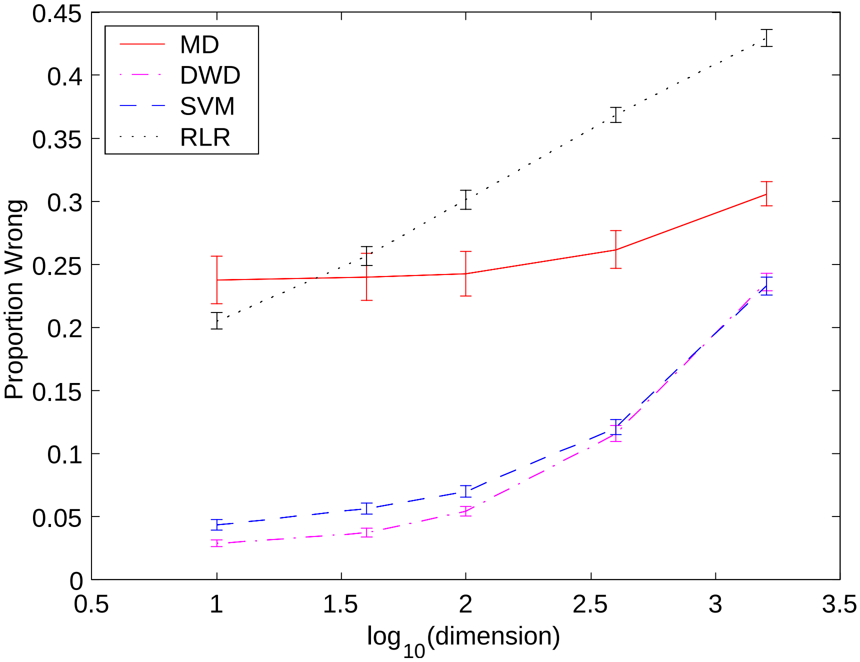

FIGURE 7: *Simulation comparison, for the wobble distribution. This is a case where DWD gives superior performance to MD and SVM.*

As in Figure 6, the few outliers have a serious and drastic effect on MD and RLR, giving them far inferior generalization performance. This time the performance of RLR is actually generally worse than MD. Note that the confidence intervals for MD are much wider, suggesting much less consistent behavior than for the other methods. Because the outliers directly impact the margin, SVM is 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 8 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 Gaussianity that drove the above examples. This time, the important method of "polynomial embedding," based on the ideas of Aizerman, Braverman and Rozoner (1964), is considered. 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 scale factor is chosen to make the "amount of separation" comparable to that in Figure 5, except that instead of "separation of the means," it is "separation in a radial direction." In particular the first $d/2$ coordinates of the data are *nested Gaussian spheres*. 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). Polynomial embedding provides a simple, yet elegant, solution to the problem of transforming the data so that separating hyperplanes become useful. In the present case, this 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.

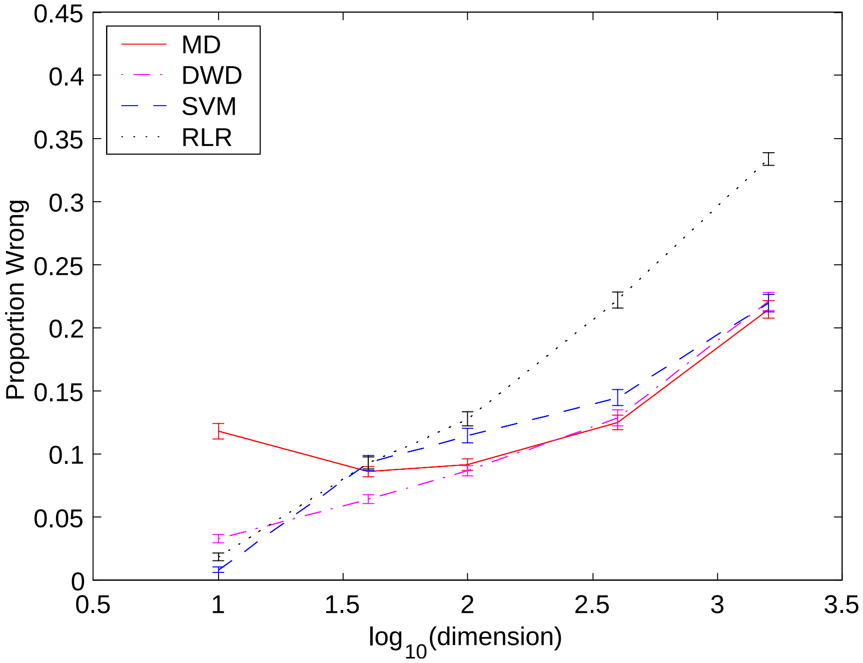

FIGURE 8: *Simulation comparison, for the nested spheres distribution. This case shows a fair overall summary, because each method is best for some $d$, and DWD tends to be near whichever method is best.*

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 $d$. This, as well as performance in some of the above examples suggests that RLR may have inferior HDLSS properties, although careful tuning (we just used simple defaults) may be able to resolve this problem.

However, because MD was motivated by Gaussian considerations (where the mean is a very important representative of the population), and the embedded data are highly non-Gaussian in nature (lying in at most a $d/2$ dimensional parabolic manifold), one might expect that MD would be somewhat inferior. However, MD is surprisingly the best of the 3 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 (except RLR) 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 (2004).

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 8 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 either MD or the SVM, in situations where it is best.

# 4 Real Data Examples

In this section we compare DWD with other methods for some real data sets. An HDLSS data set from micro-array analysis is studied in Section 4.1. A non-HDLSS data set on breast cancer diagnosis is analyzed in Section 4.2.

## 4.1 Micro-array data analysis

This section shows the effectiveness of DWD in the real data analysis of gene expression micro-array data. The data are 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. HDLSS problems are very common for micro-array data because $d$, the number of genes, can be as high as tens of thousands, and $n$, the number of cases, is usually less than 100 (often much less) because of the high cost of gathering each data point.

There are two data sets available from two studies. One is used to train the discrimination methods, and the second is used to test performance (i.e., generalizability). There are 5 classes of interest, but these are grouped into pairs because DWD is currently only implemented for 2 class discrimination. Here we consider 4 groups of pairwise problems, chosen for biological interest:

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

Group 1 Luminal cancer vs. other cancer types and normals: A first rough classification suggested by clustering of the data in Perou et al. (1999). Tested using $n_+ = 47$ and $n_- = 38$ training cases, and 51 test cases.

Group 2 Luminal A vs. Luminal B&C: an important distinction that was linked to survival rate in Perou et al. (1999). Tested using $n_+ = 35$ and $n_- = 15$ training cases, and 21 test cases.

Group 3 Normal vs. Erb & Basal cancer types. Tested using $n_+ = 13$ and $n_- = 25$ training cases, and 30 test cases.

Group 4 Erb vs. Basal cancer types. Tested using $n_+ = 11$ and $n_- = 14$ training cases, and 21 test cases.

The overall performance of the four classification methods considered in detail in this paper, over the three groups of problems, is summarized in the graphical display of Figure 9. The color of the bars indicate the classification method, and the heights show the proportion of test cases that were incorrectly classified.

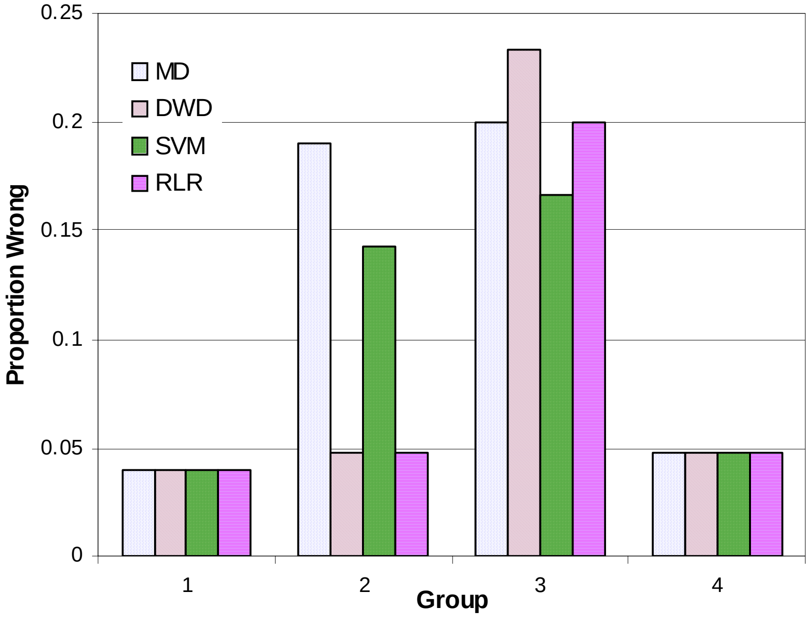

FIGURE 9: *Graphical summary of classification error rates for gene expression data.*

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 best of the others, and sometimes being better.

## 4.2 Breast Cancer Data

This section studies a much different data set, which is no longer HDLSS, the Wisconsin Diagnostic Breast Cancer data. These data were first analyzed by Street, Wolberg and Mangasarian (1993). The goal of the study was to classify $n = 569$ tumors as benign or malignant, based on $d = 30$ features that summarized available tumor information.

The same four methods as above were applied to these data. This time we study tuning parameter choice from a different viewpoint. Instead of trying to choose among the candidate tuning parameters, we study a wide range of them. This is an analog of the scale space approach to smoothing, see Chaudhuri and Marron (2000), which led to the SiZer exploratory data analysis tool proposed in Chaudhuri and Marron (1999). We suggest this approach to tuning as an interesting area for future research.

Figure 10 shows the 10-fold cross-validation scores for each method, over a wide range of tuning parameters. The SVM curve is not complete, because of computational instabilities for small values of $C$. The MD curve is constant, because this method has no tuning parameter.

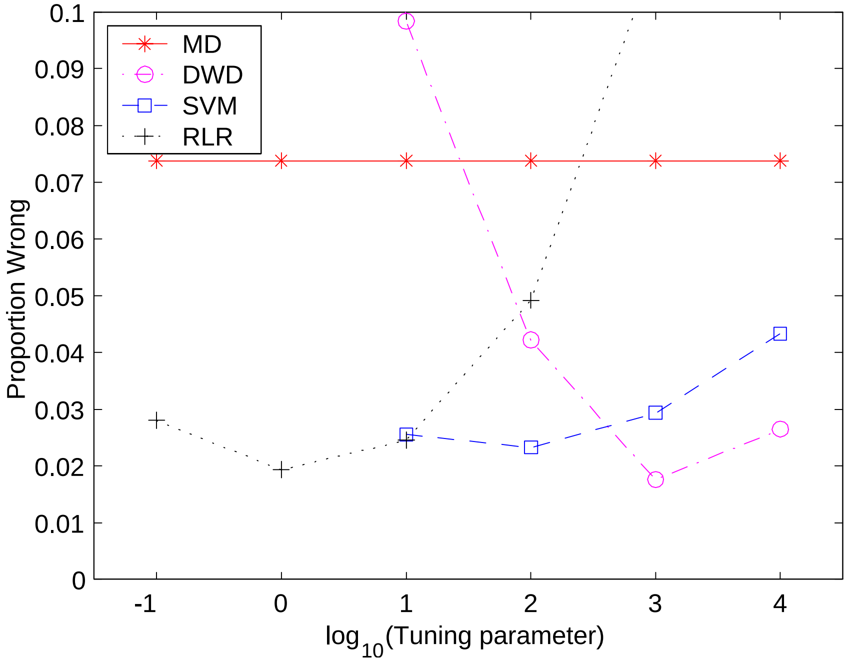

FIGURE 10: *Comparison of a range discrimination methods over a range of tuning parameters for the Wisconsin Breast Cancer data.*

As expected in this non-HDLSS setting, tuning is quite important, with each method behaving very well for some values, and quite poorly for others. Each of the tunable methods has a local minimum, which is expected for this non-HDLSS data set. DWD has the smallest minimum, but not substantially smaller than the others, so not much stock should be placed in this. A proper comparison of the values of these minima, would be via double cross-validation, where one does a cross-validated retuning of each method for each CV sub-sample, but our purpose here is just a simple scale space comparison.

The main lesson is consistent with the above: each of these methods (except perhaps MD) has the potential to be quite effective, and their relative differences are not large. Although DWD specifically targets HDLSS settings, it is good to see effective performance in other settings as well.

## 5 Open Problems

There are a number of open problems that follow from the ideas of this paper, and the DWD method.

First there are several ways in which the DWD can be fine tuned, and perhaps improved. As with the SVM, an obvious candidate for careful study is the penalty factor $C$. In many cases with separable data, the choice (if sufficiently large) will be immaterial. In a tricky case, several values of $C$ can be chosen to compare the resulting discrimination rules, but our choice provides what we believe to be a reasonable starting point. More thought could also be devoted to the choice of "typical distance" suggested in the choice of scale factor in Section 2.4. Some other movies at the web site Marron (2004) give additional insights into the effect of varying $C$. These compare DWD and SVM to the MD method, in the same toy data settings as considered in Section 1. This time data are generated in $d = 100$ and 200 dimensions, and the view is similar to the top panels of Figures 2, 3 and 4 (showing the data projected onto the plane generated by the MD direction, and either the DWD or the SVM direction). In these movies, time corresponds to changing values of $C$. There is a general tendency towards most change to happen in a relatively narrow range of $C$ values. At the upper end, this effect can be quantified by showing that both DWD and SVM are constant above a certain value of $C$. We conjecture that both methods converge to the MD in the limit as $C \to 0$. However, we encountered numerical instabilities for small values of $C$, which has limited our exploration to date in this direction. These issues will be more deeply explored in an upcoming paper. The movies in the files TwoDprojDWDd100.avi and TwoDprojDWDd200.avi are generally similar. For small values of $C$ they show that DWD is essentially the same as MD, and for large values of $C$, projections on the DWD direction, are less spread, but the subpopulations are better separated. The SVM versions of these movies, in the files TwoDprojSVMd100.avi and TwoDprojSVMd200.avi, are also quite similar to each other. But they are rather different from the corresponding DWD movies (although they were computed from the same realizations of pseudo data). For large $C$, the SVM direction exhibits very strong data piling. The data piling diminishes for smaller $C$, but some traces are still visible at all values of $C$. The SVM does not appear to converge to MD, even for the smallest $C$ values considered here. These movies show that to achieve the beneficial data separation effects of DWD, it is not enough to simply use SVM, with a lower choice of $C$.

But besides different choices of $C$, other variations that lie within the scope of SOCP optimization problems should be studied. For example, the sum of reciprocal residuals $\sum_i (1/r_i)$, could be replaced by reciprocal residuals to other powers, such as $\sum_i (1/r_i)^p$, where $p$ is a positive integer.

An important area of future research is how the separating hyperplane discrimination methods studied here can be effectively combined with dimension reduction. Bradley and Mangasarian (1998) pioneered this area with an interesting paper motivated by the SVM. An interesting improvement is the SCAD thresholding idea of Zhang et al. (2004). A major challenge is to combine *dimension reduction* (also known as *feature selection*) with DWD ideas.

Another domain of open problems is the classical statistical asymptotic analysis: When does the DWD provide a classifier that is Bayes Risk consistent? When are appropriate kernel embedded versions of either the SVM or the DWD Bayes Risk consistent? What are asymptotic rates of convergence?

Many of the issues raised in this paper can also be studied by non-standard HDLSS asymptotics, where the sample size $n$ is fixed, and the dimension $d \to \infty$. Hall, Marron and Neeman (2004) have shown that, perhaps surprisingly, such asymptotics can lead to a rigid underlying structure, which gives useful insights of a mathematical statistical nature. Much more can be done in this direction, to more deeply understand the properties of the SVM and the DWD.

Yet another domain is the performance bound approach to understanding the effectiveness of discrimination methods that has grown up in the machine learning literature. See Cannon, Ettinger, Hush and Scovel (2002), and Howse, Hush and Scovel (2002) for deep results, and some overview of this literature.

Finally, can meaningful connection between these rather divergent views of performance be established?

# 6 Appendix

This section contains details of the optimization formulations in Section 2 and their properties. Further material on the Support Vector Machine, described in Section 2.1 is in Section 6.1. A parallel detailed development of DWD, as described in Section 2.2, is given in Section 6.2.

## 6.1 SVM Optimization and its Properties

Let us first assume that $\xi = 0$. Then we can maximize the minimum $\bar{r}_i$ by solving

$$\max \quad \delta, \quad \bar{r} = YX'w + \beta y, \, \bar{r} \geq \delta e, \, w'w \leq 1,$$

where the variables are $\delta$, $\bar{r}$, $w$, and $\beta$. The constraints here are all linear except the last. Since it is easier to handle quadratics in the objective function rather than the constraints of an optimization problem, we reformulate this into the equivalent (as long as the optimal $\delta$ is positive) problem

$$\min_{w,\beta} \quad (1/2)w'w, \quad YX'w + \beta y \geq e.$$

Now we must account for the possibility that this problem is infeasible, so that nonnegative errors $\xi$ need to be introduced, with penalties; we impose a penalty on the 1-norm of $\xi$. Thus the optimization problem solved by the SVM can be stated as

$$(P_{SVM}) \qquad \min_{w,\beta,\xi} \quad (1/2)w'w + Ce'\xi, \quad YX'w + \beta y + \xi \geq e, \, \xi \geq 0.$$

where $C = C_{SVM} > 0$ is a penalty parameter.

This convex quadratic programming problem has a dual, which is

$$(D_{SVM}) \qquad \max_{\alpha} \quad -(1/2)\alpha'YX'XY\alpha + e'\alpha, \quad y'\alpha = 0, \, 0 \leq \alpha \leq Ce.$$

Further, both problems do have optimal solutions, and their optimal values are equal.

The optimality conditions for this pair of problems are:

$$\begin{array}{lll} XY\alpha = w, & y'\alpha = 0, & \\ s := YX'w + \beta y + \xi - e \geq 0, & \alpha \geq 0, & s'\alpha = 0; \\ Ce - \alpha \geq 0, & \xi \geq 0, & (Ce-\alpha)'\xi = 0. \end{array}$$

These conditions are both necessary and sufficient for optimality because the problems are convex. Moreover, the solution to the primal problem $(P_{SVM})$ is easily recovered from the solution to the dual: merely set $w = XY\alpha$ and choose $\beta = y_i - x_i'w$ for some $i$ with $0 < \alpha_i < C$. (If $\alpha = 0$, then $\xi$ must be zero and all components of $y$ must have the same sign. We then choose $\beta \in \{+1, -1\}$ to have the same sign. Finally, if each component of $\alpha$ is 0 or $C$, we can choose $\beta$ arbitrarily as long as the resulting $\xi$ is nonnegative.)

Burges (1998) notes that there is a mechanical analogy for the choice of the SVM hyperplane. Imagine that each support vector exerts a normal repulsive force on the hyperplane. When the magnitudes of these forces are suitably chosen, the hyperplane will be in equilibrium. Note that only the support vectors exert forces.

Let us give a geometrical interpretation to the dual problem, where we assume that $C$ is so large that all optimal solutions have $\alpha < Ce$. Note that $y'\alpha = 0$ implies that $e_+'\alpha_+ = e_-'\alpha_-$, where $\alpha_+$ $(\alpha_-)$ is the subvector of $\alpha$ corresponding to the Class $+1$ (Class $-1$) points and $e_+$ $(e_-)$ the corresponding vector of ones. It makes sense to scale $\alpha$ so that the sum of the positive $\alpha$'s (and that of the negative ones) equals 1; then these give convex combinations of the training points. We can write $\alpha$ in $(D_{SVM})$ as $\zeta\hat{\alpha}$, where $\zeta$ is positive and $\hat{\alpha}$ satisfies these extra scaling constraints. By maximizing over $\zeta$ for a fixed $\hat{\alpha}$, it can be seen that $(D_{SVM})$ is equivalent to maximizing $2/\|XY\hat{\alpha}\|^2$ over nonnegative $\hat{\alpha}_+$ and $\hat{\alpha}_-$ that each sum to one. But $XY\hat{\alpha} = X_+\hat{\alpha}_+ - X_-\hat{\alpha}_-$, where $X_+$ $(X_-)$ is the submatrix of $X$ corresponding to the Class $+1$ (Class $-1$) points, so we are minimizing the distance between points in the convex hulls of the Class $+1$ points and of the Class $-1$ points. Further, the optimal $w$ is the difference of such a pair of closest points.

From the optimality conditions, we may replace $w$ in $(P_{SVM})$ by $XY\alpha$, where $\alpha$ is a new unrestricted variable. Then both $(P_{SVM})$ and $(D_{SVM})$ involve the data $X$ only through the inner products of each training point with each other, given in the matrix $X'X$. This has implications in the extension of the SVM approach to the nonlinear case, where we replace the vector $x_i$ by $\Phi(x_i)$ for some possibly nonlinear mapping $\Phi$. Then we can proceed as above as long as we know the symmetric kernel function $K$ with $K(x_i, x_j) := \Phi(x_i)'\Phi(x_j)$. We replace $X'X$ with the $n \times n$ symmetric matrix $(K(x_i, x_j))$ and solve for $\alpha$ and $\beta$. We can classify any new point $x$ by the sign of

$$w'\Phi(x) + \beta = (\Phi(X)Y\alpha)'\Phi(x) + \beta = \sum_i \alpha_i y_i K(x_i, x) + \beta.$$

Here $\Phi(X)$ denotes the matrix with columns the $\Phi(x_i)$'s. It follows that knowledge of the kernel $K$ suffices to classify new points, even if $\Phi$ and thus $w$ are unknown. See Section 4 in Burges (1998).

We remark that imposing the penalty $C$ on the 1-norm of $\xi$ in $(P_{SVM})$ is related to imposing a penalty in the original maximin formulation. Suppose $w$, $\beta$, and $\xi$ solve $(P_{SVM})$ and $\alpha$ solves $(D_{SVM})$, and assume that $w$ and $\alpha$ are both nonzero. Then by examining the corresponding optimality conditions, we can show that the scaled variables $(\bar{w}, \bar{\beta}, \bar{\xi}) = (w, \beta, \xi)/\|w\|$ solve

$$\min \quad -\delta + De'\bar{\xi}, \quad YX'\bar{w} + \bar{\beta}y + \bar{\xi} \geq \delta e, \; (1/2)\bar{w}'\bar{w} \leq 1/2, \; \bar{\xi} \geq 0,$$

with $D := C/e'\alpha$. Conversely, if the optimal solution to the latter problem has $\delta$ and the Lagrange multiplier $\lambda$ for the constraint $(1/2)\bar{w}'\bar{w} \leq 1/2$ positive, then a scaled version solves $(P_{SVM})$ with $C := D/(\delta\lambda)$.

## 6.2 DWD Optimization and its Properties

Now we minimize the sum of the reciprocals of the residuals plus the parameter $C = C_{DWD} > 0$ times the 1-norm of the perturbation vector $\xi$:

$$\min_{r,w,\beta,\xi} \quad \sum_i (1/r_i) + Ce'\xi, \quad r = YX'w + \beta y + \xi, \; (1/2)w'w \leq 1/2, \; r \geq 0, \; \xi \geq 0.$$

Of course, in the problem above, $r_i$ must be positive to make the objective function finite. (More generally, we could choose the sum of $f(r_i)$'s, where $f$ is any smooth convex function that tends to $+\infty$ as its argument approaches 0 from above. However, the reciprocal leads to a nice optimization problem, as we show below.) Here we show how this can be reformulated as a SOCP problem, state its dual and the corresponding optimality conditions, discuss a dimension-reduction technique for the case that $d \gg n$, and give a geometric interpretation of the dual. We also show how the method can be extended to handle the nonlinear case using a kernel function $K$. Finally, we discuss an interpretation of the penalty parameter $C$. Further details of the arguments can be found in the paper DWD1.pdf in the web directory Marron (2004).

We first show how the reciprocals can be eliminated by using second-order cone constraints. Recall that second-order cones have the form

$$S_{m+1} := \{(\zeta; u) \in \Re^{m+1} : \zeta \geq \|u\|\}.$$

To do this, write $r_i = \rho_i - \sigma_i$, where $\rho_i = (r_i + 1/r_i)/2$, $\sigma_i = (1/r_i - r_i)/2$. Then $\rho_i^2 - \sigma_i^2 = 1$, or $(\rho_i; \sigma_i; 1) \in S_3$, and $\rho_i + \sigma_i = 1/r_i$. We also write $(1/2)w'w \leq 1/2$ as $(1; w) \in S_{d+1}$. We then obtain

$$\begin{array}{lcccccccccccc}
\min_{\psi,w,\beta,\xi,\rho,\sigma,\tau} & & & & & & Ce'\xi & + & e'\rho & + & e'\sigma & & \\
 & & YX'w & + & \beta y & + & \xi & - & \rho & + & \sigma & = & 0, \\
 & \psi & & & & & & & & & & = & 1, \\
(P_{DWD}) & & & & & & & & & & \tau & = & e,
\end{array}$$

$$(\psi; w) \in S_{d+1}, \quad \xi \geq 0, \quad (\rho_i; \sigma_i; \tau_i) \in S_3, \; i = 1, 2, \ldots, n.$$

SOCP problems have nice duals. Indeed, after applying the usual rules for obtaining the dual and simplifying the result, we obtain

$$(D_{DWD}) \qquad \max_{\alpha} \quad -\|XY\alpha\| + 2e'\sqrt{\alpha}, \quad y'\alpha = 0, \; 0 \leq \alpha \leq Ce.$$

This is very similar to $(D_{SVM})$ above, which instead maximizes $-(1/2)\|XY\alpha\|^2 + e'\alpha$.

It is straightforward to show that the sufficient conditions for existence of optimal solutions and equality of the optimal objective function values of $(P_{DWD})$ and $(D_{DWD})$ hold. See again the paper cited above. Further it can be shown that the optimality conditions can be written as

$$\begin{array}{ll} YX'w + \beta y + \xi - \rho + \sigma = 0, & y'\alpha = 0, \\ \alpha > 0, \quad \alpha \leq Ce, \quad \xi \geq 0, & (Ce - \alpha)'\xi = 0, \\ \text{Either } XY\alpha = 0 & \text{ and } \|w\| \leq 1, \\ \text{or } w = XY\alpha/\|XY\alpha\|, & \\ \rho_i = (\alpha_i + 1)/(2\sqrt{\alpha_i}), & \sigma_i = (\alpha_i - 1)/(2\sqrt{\alpha_i}), \text{ for all } i. \end{array}$$

In the HDLSS setting, it may be inefficient to solve $(P_{DWD})$ and $(D_{DWD})$ directly. Indeed, the primal variable $w$ is of dimension $d \gg n$, the sample size. Instead, we can proceed as follows. First factor $X$ as $QR$, where $Q \in \Re^{d\times n}$ has orthonormal columns and $R \in \Re^{n\times n}$ is upper triangular: this can be done by a (modified) Gram-Schmidt procedure or by orthogonal triangularization, see, e.g., Golub and Van Loan [15]. Then we can solve $(P_{DWD})$ and $(D_{DWD})$ with $X$ replaced by $R$, so that in the primal problem $YR'\bar{w}$, with $(\psi; \bar{w}) \in S_{n+1}$, replaces $YX'w$, with $(\psi; w) \in S_{d+1}$. Thus the number of variables and constraints depends only on $n$, not $d$.

Note that, since $X' = R'Q'$, any feasible solution $(\psi, \bar{w}, \beta, \xi, \rho, \sigma, \tau)$ of the new problem gives a feasible solution $(\psi, w, \beta, \xi, \rho, \sigma, \tau)$ of the original problem on setting $w = Q\bar{w}$ ($\|w\| = \|\bar{w}\|$), since $YX'w = YR'Q'Q\bar{w} = YR'\bar{w}$; moreover, this solution has the same objective value. We therefore solve the new smaller problems and set $w = Q\bar{w}$ to get an optimal solution to the original problem. (We can also avoid forming $Q$, even in product form [15], finding $R$ by performing a Cholesky factorization $R'R = X'X$ of $X'X$; if $R$ is nonsingular, we recover $w$ as $XR^{-1}\bar{w}$, but the procedure is more complicated if $R$ is singular, and we omit details.)

There is again a mechanical analogy for the separating hyperplane found by the DWD approach (we assume that all optimal solutions to $(D_{DWD})$ have $\alpha < Ce$). Indeed, the function $1/r$ is the potential for the force $1/r^2$, so the hyperplane is in equilibrium if it is acted on by normal repulsive forces with magnitude $1/r_i^2$ at each training point. Indeed, $r_i = \rho_i - \sigma_i = 1/\sqrt{\alpha_i}$ at optimality, so the force is $\alpha_i$ at training point $x_i$. The dual constraint $y'\alpha = 0$ implies that the vector sum of these forces vanishes, and the fact that $XY\alpha$ is proportional to $w$ from the optimality conditions implies that there is no net torque either.

We now give an interpretation of $(D_{DWD})$, similar to that of finding the closest points in the two convex hulls for $(D_{SVM})$. Indeed, if we again write $\alpha$ as $\zeta\hat{\alpha}$, where $\zeta$ is positive and $e'_+\hat{\alpha}_+ = e'_-\hat{\alpha}_- = 1$, the objective function becomes

$$\max_{\zeta,\hat{\alpha}} -\zeta\|XY\hat{\alpha}\| + 2\sqrt{\zeta}e'\sqrt{\hat{\alpha}},$$

and if we maximize over $\zeta$ for fixed $\hat{\alpha}$ we find $\zeta = (e'\sqrt{\hat{\alpha}}/\|XY\hat{\alpha}\|)^2$. Substituting this value, we see that we need to choose convex weights $\hat{\alpha}_+$ and $\hat{\alpha}_-$ to maximize

$$(e'_+\sqrt{\hat{\alpha}_+} + e'_-\sqrt{\hat{\alpha}_-})^2/\|X_+\hat{\alpha}_+ - X_-\hat{\alpha}_-\|.$$

Thus we again want to minimize the distance between points in the two convex hulls, but now divided by the square of the sum of the square roots of the convex weights. This puts a positive weight on every training point. As long as the convex hulls are disjoint, the difference of these two points, $XY\hat{\alpha} = X_+\hat{\alpha}_+ - X_-\hat{\alpha}_-$, will be nonzero, and the normal to the separating hyperplane will be proportional to this vector by the optimality conditions.

In the case that we expect, $XY\alpha \neq 0$, $w$ has the form $XY\alpha$ for a (scaled) $\alpha$. Hence it seems that we can once again handle the nonlinear case using a kernel function $K$. But software for SOCP problems assumes the formulation is as above, i.e., we cannot replace $w$ by $XY\alpha$ and add the constraint $\alpha YX'XY\alpha \leq 1$. Instead we can proceed as follows. Indeed, this approach also works in the exceptional case, as we see below.

Form the matrix $M := (K(x_i, x_j))$ as in the SVM case, and factorize it as $M = R'R$, e.g., using the Cholesky factorization. Now replace $YX'w$ by $YR'\bar{w}$ in $(P_{DWD})$, and replace $(\psi; w) \in S_{d+1}$ by $(\psi; \bar{w}) \in S_{n+1}$. Similarly, in $(D_{DWD})$ replace $XY\alpha$ by $RY\alpha$. (This is like the dimension-reducing technique discussed above.) Suppose we solve the resulting problems to get $\bar{w}$, $\alpha$ and $\beta$.

If $RY\alpha \neq 0$, then it follows as in the linear case that $\bar{w} = RY\bar{\alpha}$, where $\bar{\alpha} := \alpha/\|RY\alpha\|$. But even if $RY\alpha = 0$, we note that $\bar{w}$ appears in $(P_{DWD})$ only in the constraints $YR'\bar{w} + \cdots = 0$ and $(\psi; \bar{w}) \in S_{n+1}$, and so we can replace $\bar{w}$ by the minimum norm $\hat{w}$ with $YR'\hat{w} = YR'\bar{w}$. The optimality conditions of this linear least-squares problem imply that $\hat{w} = RY\bar{\alpha}$ for some $\bar{\alpha}$. We claim that we can classify a new point $x$ as before by the sign of $\sum_i \bar{\alpha}_i y_i K(x_i, x) + \beta$, where $\bar{\alpha}$ is obtained by one of the two methods above.

Indeed, since we can restrict $\bar{w}$ to be of the form $RY\bar{\alpha}$, $(P_{DWD})$ is equivalent to the problem with $YR'\bar{w}$ replaced by $YR'RY\bar{\alpha}$, and $(\psi; \bar{w}) \in S_{n+1}$ replaced by $\bar{\alpha}'YR'RY\bar{\alpha} \leq 1$; $\bar{w}$ can then be retrieved by setting it to $RY\bar{\alpha}$. Now we can make the same argument for the version of $(P_{DWD})$ with $Y\Phi(X)'w + \cdots = 0$ and $(\psi; w) \in S_{d+1}$. We can assume that $w$ is of the form $\Phi(X)Y\tilde{\alpha}$ and substitute for $w$ to get $Y\Phi(X)'\Phi(X)Y\tilde{\alpha} + \cdots = 0$ and $\tilde{\alpha}'Y\Phi(X)'\Phi(X)Y\tilde{\alpha} \leq 1$, and then recover $w$ as $\Phi(X)Y\tilde{\alpha}$. But the two problems, one with $\bar{\alpha}$ and one with $\tilde{\alpha}$, are identical, since

$$YR'RY = YMY = Y(K(x_i, x_j))Y = Y\Phi(X)'\Phi(X)Y,$$

and so both have identical optimal solutions, and hence we can classify new points by the sign of $w'\Phi(x) + \beta = \bar{\alpha}'Y\Phi(X)'\Phi(x) + \beta = \sum_i \bar{\alpha}_i y_i K(x_i, x) + \beta$, as claimed.

We should note that the "bad" case $XY\alpha = 0$ can happen, e.g., with $n = 2$, $x_1 = x_2$, and $y_1 = -y_2$. Then $\alpha_1 = \alpha_2 = C$ and $XY\alpha = 0$. But in this case, all we need is the extra solution of a linear least-squares problem.

Let us give an interpretation of the penalty parameter $C = C_{DWD}$ and some suggestions on how it can be set. Recall that $\bar{r}$ is the unperturbed residual, so that $\bar{r}_i := y_i(w'x_i + \beta)$, which can be of any sign. If this quantity is given, $(P_{DWD})$ will choose the nonnegative perturbation $\xi_i$ to minimize $1/(\bar{r}_i + \xi_i) + C\xi_i$. It is easy to see that the resulting $\xi_i$ (if positive) satisfies $(\bar{r}_i + \xi_i)^{-2} = C$, so that $\bar{r}_i + \xi_i = C^{-1/2}$. This is the argument where the derivative of the function $f(t) := 1/t$ has slope $-C$, and it is not hard to check that the contribution of the $i$th data point to the objective function of $(P_{DWD})$ is $\bar{f}(\bar{r}_i)$, where $\bar{f}$ is the function that agrees with $f$ to the right of $C^{-1/2}$, and is a straight line with slope $-C$ to the left, with the constant part chosen to make the function continuous (and continuously differentiable). Hence instead of perturbing the residuals and penalizing the amount of perturbation, we can view the approach as perturbing the criterion function $f$ so that it applies to negative as well positive residuals.

This suggests using a value for $C$ that is a typical slope of the reciprocal function. Hence we find that $C$ should scale with the inverse square of a distance between the training points, but not with the number of training points, and similarly to the SVM case, a reasonable value will be a large constant divided by the square of a typical distance between training points. This recommendation is also suggested by examining the optimality conditions, because if all $x_i$'s are scaled by $\gamma$, we expect $w$ to be the same (since in the usual case its norm is one), so that $\beta$, $\xi$, and $\rho - \sigma$ are scaled by $\gamma$. Then $\alpha$ is scaled by $\gamma^{-2}$, and so $C$ is scaled by the same factor. If training points are replicated, then the vectors $\xi$, $\rho$, $\sigma$, and $\alpha$ are similarly expanded, but their components, and hence $C$, remain the same.

SOCP problems are certainly much less well-known in optimization than quadratic programming problems as in the SVM approach. However, there has been rising interest in them recently, because of their power in modeling and their amenability to efficient algorithms. See Alizadeh and Goldfarb (2003), Lobo, Vandenberghe, Boyd and Lebret (1998), Nesterov and Todd (1997, 1998), Tsuchiya (1999) and Tütüncü, Toh, and Todd (2001b). We used the SDPT3 package described in the last paper.

# 7 Acknowledgement

The research of J. S. Marron was supported by Cornell University's College of Engineering Mary Upson Fund and NSF Grants DMS-9971649 and DMS-0308331. M. J. Todd was supported in part by NSF Grants DMS-9805602 and DMS-0209457 and ONR Grant N00014-02-1-0057. Marron was grateful for the chance to spend a year in the exciting research environment of the School of Operations Research and Industrial Engineering, from which this collaboration is a direct result.

# References

[1] Ahn, J. and Marron, J. S. (2004) Maximal data piling in discrimination, manuscript in preparation.

[2] 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.

[3] Alizadeh, F. and Goldfarb, D. (2003) Second-Order Cone Programming, *Mathematical Programming*, 95, 3–51. Available at http://rutcor.rutgers.edu/~rrr/2001.html.

[4] 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.

[5] Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's Linear Discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations." U of Michigan, Dept of Statistics Technical Report #404, (to appear in *Bernoulli*).

[6] Bradley, P. S. and Mangasarian, O. L. (1998) Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference (ICML '98)*, 82–90, Morgan Kaufmann, San Francisco. Available at ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98–03.ps.

[7] Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2, 955–974. Available at http://citeseer.nj.nec.com/burges98tutorial.html.

[8] Cannon, A., Ettinger, J. M., Hush, D. and Scovel, C. (2002) Machine learning with data dependent hypothesis classes, *Journal of Machine Learning Research*, 2, 335–358.

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

[10] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807–823.

[11] Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation, *Annals of Statistics*, 28, 408–428.

[12] Cover, T. (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Transactions on Electronic Computers*, 12, 326–334, reprinted in (1992) *Artificial Neural Networks: Concepts and Theory*, eds. P. Mehra and B. Wah, IEEE Computer Society Press, Los Alamitos, Calif.

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

[14] Duda, R. O., Hart, P. E. and Stork, D. G. (2000) *Pattern Classification*, John Wiley & Sons, New York.

[15] Golub, G. H., and Van Loan, C. F. (1989) *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD.

[16] Hall, P., Marron, J. S. and Neeman, A. (2004) Geometric representation of high dimension low sample size data, submitted to *Journal of the Royal Statistical Society, Series B.*

[17] Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*, Springer Verlag, Berlin.

[18] Howse, J., Hush, D. and Scovel, C. (2002) Linking learning strategies and performance for support vector machines, *Los Alamos National Laboratory Technical Report LA-UR-02-1933.*

[19] Koch, I., Marron, J. S. and Chen, Z. (2004) Simulation of non-Gaussian populations of images, unpublished manuscript.

[20] Lin, Y., Wahba, G., Zhang, H., and Lee, Y. (2002) Statistical properties and adaptive tuning of support vector machines, *Machine Learning*, 48, 115–136, 2002. Available at ftp://ftp.stat.wisc.edu/pub/wahba/index.html.

[21] Lobo, M. S., Vandenberghe, L., Boyd, S. and Lebret, H. (1998) Applications of second-order cone programming, *Linear Algebra and Its Applications*, 284, 193–228.

[22] Marron, J. S. (2004) Web site:
http://www.stat.unc.edu/postscript/papers/marron/HDD/DWD/.

[23] Marron, J. S., Wendelberger, J. and Kober, E. (2004) Time series functional data analysis, unpublished manuscript.

[24] Nesterov, Yu. E. and Todd, M. J. (1997) Self-scaled barriers and interior-point methods for convex programming, *Mathematics of Operations Research*, 22, 1–42.

[25] Nesterov, Yu. E. and Todd, M. J. (1998) Primal-dual interior-point methods for self-scaled cones, *SIAM Journal on Optimization*, 8, 324–364.

[26] Peng, F. and McCallum, A. (2004) Accurate information extraction from research papers using conditional random fields, *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2004.

[27] 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. Web site: http://genome-www.stanford.edu/sbcmp/index.shtml

[28] Schimek, M. (2003) Penalized logistic regression in gene expression analysis, *Proceedings to 2003 Semiparametric Conference*, Berlin, Germany.

[29] Schölkopf, B. and Smola, A. J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, The MIT Press, Cambridge, Massachusetts.

[30] 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. See also http://www.cs.wisc.edu/~olvi/uwmp/cancer.html.

[31] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Science*, 99, 6567–6572.

[32] Tsuchiya, T. (1999) A convergence analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming, *Optimization Methods and Software*, 11/12, 141–182.

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

[34] Tütüncü, R. H., Toh, K. C., and Todd, M. J. (2001b) SDPT3 – a MATLAB software package for semidefinite-quadratic-linear programming, available from http://www.math.cmu.edu/ reha/home.html (August 2001).

[35] Vapnik, V. N. (1982) *Estimation of Dependences Based on Empirical Data*, Springer Verlag, Berlin (Russian version, 1979).

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

[37] Wahba, G., Lin, Y., Lee, Y. and Zhang, H. (2001) Optimal properties and adaptive tuning of standard and nonstandard support vector machines, to appear in *Proceedings of the MSRI Berkeley Workshop on Nonlinear Estimation and Classification.* Available at ftp://ftp.stat.wisc.edu/pub/wahba/index.html.

[38] Wright, S. J. (1997) *Primal-dual Interior-point Methods*, SIAM, Philadelphia.

[39] Yushkevich, P., Pizer, S. M., Joshi, S. and Marron, J. S. (2001) Intuitive, Localized Analysis of Shape Variability, in *Information Processing in Medical Imaging (IPMI)*, eds. Insana, M. F. and Leahy, R. M., 402–408.

[40] Zhang, H., Ahn, J., Lin, X. and Park, C. (2004) Variable selection for SVM using nonconcave penalty, unpublished manuscript.

[41] Zhang, T. and Oles, F. (2001) Text classification based on regularized linear classifiers, *Information Retrieval*, 4, 5–31.