Robust Centroid Quantile Based Classification for High Dimension Low Sample Size Data

Jiancheng Jiang
Department of Probability and Statistics
Peking University
Beijing 100871, China

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260 USA

February 17, 2003

Abstract

A new method of statistical classification (discrimination) is proposed. The method is most effective for high dimension low sample size data. Its value is demonstrated through a new type of asymptotic analysis, and via a simulation study.

Keywords: Classification, high dimension, low sample size, quantile, robust centroid.
Footnote: Jiancheng Jiang was supported by Chinese NSF Grant 10001004.

1 Introduction

High dimension, low sample size (HDLSS) data present special challenges to many classical statistical techniques. For example, in much of standard multivariate analysis, the usual first step of "sphering the data", by multiplying by the root inverse covariance matrix, is impossible because the empirical covariance matrix is not of full rank. Because such data are appearing more and more frequently in a variety of settings, including micro-arrays for gene expression, medical image analysis, and chemometrics, it is no longer appropriate to simply insist that more data must be gathered before analysis. This motivates a need for development of a large range of new (multivariate) statistical procedures.

Here the statistical context of classification (also called discrimination) is considered. The common approaches of Fisher Linear Discrimination (FLD) and Gaussian Likelihood ratio are generally inadequate. The Gaussian Likelihood ratio fails completely in the HDLSS case, because there is no density available. The standard FLD formula cannot be computed because the empirical covariance matrix is not invertible. This hurdle can be overcome,

using a generalized inverse, but the result is usually poor, because a spurious direction is usually found. In particular, when the dimensionality is higher than the sample size, FLD will often find a direction vector with the property that the projected data for each class will pile up on a single point. Such a direction is "perfect" for discrimination of the given data, but is usually very poor for new data, i.e. it results in poor "generalization performance".

A very simple and intuitive approach to discrimination in HDLSS situations is the "Mean Difference" (MD) method, where one simply uses the direction vector which is the difference of the two mean vectors. A new data point is projected onto the difference vector, and the class whose mean is closest to the given data point is chosen. A shrinkage based refinement of this idea gives the "centroid method" of Tibshirani, Hastie, Narasimhan and Chu (2002a,b). This method is Bayes Risk Optimal, regardless of the dimension, if both class distributions are multivariate Gaussian, with the same spherical covariance structure. However, in other cases, either non-Gaussian distributions, or non-identical, or non-spherical covariance, it can be far from effective.

This motivates a search for improved methods of classification in HDLSS settings. The Support Vector Machine (SVM), see Vapnik (1982,1995), has improved properties of this type. However, as noted by Marron and Todd (2002), this has some undesirable properties for HDLSS situations, in terms of data (projected onto the direction vector) also piling at the margin. Marron and Todd went on to propose Distance Weighted Discrimination (DWD), which like the SVM relies on sophisticated optimization techniques. DWD gives superior performance to SVM for HDLSS data because it replaces the margin based optimization criterion with a "distance weighted" version, which avoids the data piling on the margin.

In this paper, we propose a simpler method, that also gives effective performance in a variety of HDLSS situations. Starting with the Mean Difference idea, we provide robustness against non-Gaussian distributions by replacing the sample means with Huber's L^1 M-estimate, a much more robust notion of "center". We also address the problem of non-spherical covariance, by using a rank based quantile method, on the data projected onto the difference vector, for deciding the final classification. The resulting method is called Robust Centroid Quantile (RCQ) classification.

Details of the RCQ discrimination method are given in Section 2. The use of asymptotic analysis for assessment and comparison to other methods can be found in Section 3. The asymptotics are along the lines of Hall and Marron (2003), and are completely different from those appearing elsewhere (yet are appropriate for HDLSS settings) in that the sample size is fixed, and the dimension tends to infinity. This mode of asymptotics results in a particular limiting geometry, that is explained in Section 3.1. This geometry is then used to understand how SVM, DWD and RCQ relate to each other in Section 3.2, where it is seen that the direction chosen by RCQ has less variability than SVM direction (a property shared by DWD), and that RCQ has better properties in the unequal sample size case than DWD. Further investigation of this comparison is done in a simulation study in Section 4, where it is seen that as expected, no method is uniformly best. RCQ is the best, or else quite similar to the best, rather often. It is especially strong in situations with non-spherical covariance structure.

2 RCQ Classification

Suppose that $\mathcal{X} = \{X_1, \cdots, X_{n_1}\}$ and $\mathcal{Y} = \{Y_1, \cdots, Y_{n_2}\}$, are two independent training sets, which are each iid samples from d-dimensional populations. In the HDLSS case, $n_1, n_2 \ll d$. For medical image analysis examples, n_1 and n_2 are frequently in the range of 20 to 100, and d is in the high tens to hundreds. In the case of micro-array gene expression data, n_1 and n_2 are usually in the lower tens, while d is in the thousands to tens of thousands. Many chemical spectral data sets have n_1 and n_2 also in the tens, and d in the thousands.

The classification (discrimination) problem is to assign new observations, i.e. new d-dimensional vectors to Class $\mathcal{X}$ or Class $\mathcal{Y}$, depending on which is "most appropriate". See Duda, Hart and Stork (2000) for an overview of available classification methods. In this paper, we restrict attention to two class methods for simplicity. We also restrict attention to methods which have give a "direction vector" where projection of the data gives effective separation of the classes. We prefer such methods, because they frequently give insight into differences between the classes, of a type that is unavailable from the widely studied nearest neighbor and neural network methods.

As noted in Section 1, the SVM is a promising method of this type, which is substantially improved by DWD in HDLSS settings. In this paper, we propose a classification rule which is simpler than DWD, based on robust centroid fitting and quantile searching.

Let C_X denote a robust centroid of the $\mathcal{X}$ population, and C_Y that of the $\mathcal{Y}$ population. There are many choices of "robust population centroid", and a large literature on this topic, see e.g. Hampel, Ronchetti, Rousseeuw and Stahel (1986), Huber (1981), Rousseeuw and Leroy (1987) and Staudte and Sheather (1990).

The new RCQ method proposed in this paper proceeds as follows:

(i) Find robust estimators $\hat{C}_X$ and $\hat{C}_Y$ of C_X and C_X respectively. We take $\hat{C}_X$ and $\hat{C}_Y$ to be "L^1 M-estimate of location", detailed in Section 2.1, although many other choices are possible as well. Let $\hat{C}_{XY}$ be the vector $\hat{C}_Y - \hat{C}_X$ and let $\mathcal{X}^* = \{X_1^*, \cdots, X_{n_1}^*\}$ denote the projection of $\mathcal{X}$ onto the unit vector $\hat{C}_{XY} / \left\| \hat{C}_{XY} \right\|$, and let $\mathcal{Y}^* = \{Y_1^*, \cdots, Y_{n_2}^*\}$ denote the same projections of $\mathcal{Y}$.

(ii) The classification boundary in $\mathbb{R}^d$ is the hyperplane $\mathbf{P}$ (the *RCQ* plane) whose unit normal vector is $\hat{C}_{XY} / \left\| \hat{C}_{XY} \right\|$, and whose intercept is defined in terms of the projected data:

$$\hat{C}^* = \text{median} \left\{ C : \ \frac{1}{n_1} \sum_{i=1}^{n_1} 1(X_i^* \leq C) = \frac{1}{n_2} \sum_{j=1}^{n_2} 1(Y_j^* \geq C) \right\}. \tag{1}$$

(iii) Classify a new vector Z as coming from the $X-$ or $Y-$ population according to its position with respect to the hyperplane $\mathbf{P}$. In particular, assign Z to Class $\mathcal{X}$ when the inner product of Z with $\hat{C}_{XY} / \left\| \hat{C}_{XY} \right\|$ is $\leq \hat{C}^*$.

Since the solution set of the inner equality in (1) may be an either a point or an interval, we take $\hat{C}^*$ as the median to make it "most representative".

Note that $\hat{C}^*$ can be viewed as an empirical estimator of the population version

$$C^* = \text{median}\{C: \ P(X^* \leq C) = P(Y^* \geq C)\}, \tag{2}$$

where X^* and Y^* represent projection onto the population normal vector $C_{XY} \equiv C_Y - C_X$. The discrimination rule, including the cutoff point C^* and the hyperplane $\mathbf{P}$ can be interpreted in terms of $\mathbb{R}^d$ without projecting onto C_{XY}.

The cutoff $\hat{C}^*$, defined at (1) can be computed as

$$\hat{C}^* \equiv \text{median}\{c: \ F_X^*(c) = 1 - G_Y^*(c)\}, \tag{3}$$

where F_X^* and G_Y^* are empirical distribution functions of X^* and Y^*, respectively. We approximated these functions, by linear interpolation, over an equally spaced grid of 100 points from the median of $\mathcal{X}^*$ to the median of $\mathcal{Y}^*$.

2.1 Robust Centroid Estimation

The robust centroid estimate studied here is the "L^1 M-estimate of location", see Section 6.3 of Huber (1981). Given a multivariate data set, such as $\mathcal{X} = \{X_1, \cdots, X_{n_1}\} \subseteq \mathbb{R}^d$, this is defined as:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n_1} \|X_i - \theta\|_2 ,$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm on $\mathbb{R}^d$. Note that $\widehat{\theta}$ may be found as the solution of the equation:

$$0 = \frac{\partial}{\partial\theta} \sum_{i=1}^{n_1} \|X_i - \theta\|_2^p = \sum_{i=1}^{n_1} \frac{X_i - \theta}{\|X_i - \theta\|_2}. \tag{4}$$

Insight as to how this location estimate dampens the effect of outliers comes from recognizing that

$$\frac{X_i - \theta}{\|X_i - \theta\|_2} + \theta = P_{Sph(\theta,1)} X_i,$$

i.e. the projection of X_i onto the sphere centered at θ, with radius 1. Thus the solution of (4) is the solution of

$$0 = avg\left\{P_{Sph(\theta,1)} X_i - \theta : i = 1, ..., n_1\right\}.$$

Hence $\widehat{\theta}$ may be understood by considering candidate unit spheres centered at θ, projecting the data onto the sphere, then moving the sphere around until the average of the projected values is at the center of the sphere. These ideas are illustrated in Figure 4.2 of Locantore, et. al (1999).

It can be shown that in one dimension, $\widehat{\theta}$ is any sample median. Hence $\widehat{\theta}$ has been called "the spatial median" for higher dimensions. Another consequence is that this location estimate is not unique. However, Milasevic and Ducharme (1987) have shown that in more than one dimension, $\widehat{\theta}$ is unique, unless all of the data lie in a one dimensional subspace. Other terminology has also been used, e.g. Haldane (1948) called it the "geometric median" and made very early remarks on its robustness properties.

A simple and direct iterative method for calculating $\widehat{\theta}$ comes from Gower (1974) or from Section 3.2 of Huber (1981). Given an initial guess, $\widehat{\theta}_0$, iteratively define:

$$\widehat{\theta}_\ell = \frac{\sum_{i=1}^{n_1} w_i X_i}{\sum_{i=1}^{n_1} w_i}$$

where

$$w_i = \frac{1}{\left\| X_i - \widehat{\theta}_{\ell-1} \right\|_2}.$$

This iteration can be understood through the relationship

$$\widehat{\theta}_\ell = \widehat{\theta}_{\ell-1} + \frac{\sum_{i=1}^{n_1} w_i \left(X_i - \widehat{\theta}_{\ell-1} \right)}{\sum_{i=1}^{n_1} w_i} = \widehat{\theta}_{\ell-1} + \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} P_{Sph(\theta_{\ell-1},1)} X_i - \widehat{\theta}_{\ell-1}}{\frac{1}{n_1}\sum_{i=1}^{n_1} w_i}.$$

This shows that the next step is in the direction of the vector from the current sphere center $\widehat{\theta}_{\ell-1}$ to the mean of the projected data, $\frac{1}{n_1}\sum_{i=1}^{n_1} P_{Sph(\theta_{\ell-1},1)} X_i$. The length of the step is weighted by the harmonic mean distance of the original data to the sphere center (so larger steps are taken when the data are more spread). We take $\widehat{\theta}_0$ to be the sample mean, and iterate until either 20 steps have been taken, or the relative difference between $\widehat{\theta}_\ell$ and $\widehat{\theta}_{\ell-1}$ was less than 10^{-6}. More work should be done on verification and fine tuning of these choices in HDLSS setting, and it may be useful to use a different starting point, such as the coordinate-wise median.

3 Properties

In this section, asymptotic analysis is done to compare the classification methods RCQ, SVM and DWD. The asymptotics are unusual, and follow the ideas of Hall and Marron (2003), who discovered an interesting geometry, discussed in Section 3.1.

This asymptotic analysis studies the limit as $d \to \infty$, for fixed n_1 and n_2. Letting $X^{(k)}$ denote the k-th entry of the vector X, useful technical assumptions are:

On average, the variance of the entries of the Class $\mathcal{X}$ data vectors is σ^2:

$$\frac{1}{d}\sum_{k=1}^{d} var(X^{(k)}) \to \sigma^2. \tag{5}$$

On average, the variance of the entries of the Class $\mathcal{Y}$ data vectors is τ^2:

$$\frac{1}{d}\sum_{k=1}^{d} var(Y^{(k)}) \to \tau^2. \tag{6}$$

The average squared difference between means of the entries of the Class $\mathcal{X}$ and Class $\mathcal{Y}$ data vectors is μ^2:

$$\frac{1}{d}\sum_{k=1}^{d} [E(X^{(k)} - E(Y^{(k)}))]^2 \to \mu^2. \tag{7}$$

$$X \text{ and } Y \text{ are independent.} \tag{8}$$

Laws of Large Numbers for the sequence of entries in the Class $\mathcal{X}$ and Class $\mathcal{Y}$ data vectors follow from:

$$\text{The components of } X \text{ and } Y \text{ form } \rho \text{ mixing time series,} \tag{9}$$

and from the moment conditions:

$$\sup_{k=1,\ldots,d} E\left(X^{(k)}\right)^4 \leq M, \tag{10}$$

$$\sup_{k=1,\ldots,d} E\left(Y^{(k)}\right)^4 \leq M, \tag{11}$$

for some M.

3.1 Geometry

Assume that the conditions above hold. From the geometric representation, developed by Hall and Marron (2003) for the points in the sample, $\mathcal{X}$ say, we know that the distance between X_i and X_j, for any $i \neq j$, approximately equals $\sqrt{2\sigma^2 d}$ as $d \to \infty$, in the sense that

$$\frac{1}{\sqrt{d}}\{\sum_{k=1}^{d}(X_i^{(k)} - X_j^{(k)})^2\}^{\frac{1}{2}} \xrightarrow{p} \sqrt{2\sigma^2},$$

as $d \to \infty$, where $X_i^{(k)}$ is the k-th component of the vector X_i. Similarly, the distance between Y_i and Y_j, for any $i \neq j$, approximately equals $\sqrt{2\tau^2 d}$, and the distance between X_i and Y_j, for any $i,\ j$, approximately equals $\ell \equiv \sqrt{(\sigma^2 + \tau^2 + \mu^2)d}$ as $d \to \infty$. Then after rescaling each component of d-variate space by the factor $d^{-1/2}$, we have the following geometric structure for the samples $\mathcal{X}$ and $\mathcal{Y}$ (see Section 3.2 of Hall and Marron 2003):

> *After rescaling each component of d-variate space by the factor $d^{-1/2}$, the $N = n_1 + n_2$ points in $\mathcal{X} \cup \mathcal{Y}$ are asymptotically located at the vertices of a convex N-polyhedron in $(N-1)$-dimensional space, where the polyhedron has N vertices and $N(N-1)/2$ edges. Just n_1 of the vertices are the limits of the n_1 points of $\mathcal{X}$, and are the vertices of an n_1-simplex of edge length $2^{1/2}\sigma$. The other n_2 vertices are the limits of the n_2 points of $\mathcal{Y}$, and are the vertices of an n_2-simplex of edge length $2^{1/2}\tau$. The lengths of the edges in the N-polyhedron that link a vertex deriving from a point in $\mathcal{X}$ to one deriving from a point in $\mathcal{Y}$, are all of length $\tilde{}\ell$. An N-polyhedron is a figure in $(N-1)$-dimensional space that has just N vertices and has all its faces given by planes in $(N-1)$-variate space. The particular one discussed above has all the scale-invariant properties of an N-simplex, and in particular has just $\binom{N}{k}$ k-faces, or faces that are of dimension $k-1$. Thus, it has $\binom{N}{1}$ vertices, $\binom{N}{2}$ edges, and so on.*

The SVM method chooses the hyperplane which perpendicularly bisects the two closest points in the convex hulls of the respective datasets. Asymptotically, the convex hulls are precisely the n_1- and n_2- simplices the vertices of which represent the limits, as $d \to \infty$,

of the datasets$\mathcal{X}$ and $\mathcal{Y}$, respectively (see Hall and Marron, 2003). Note that each pair of points from $\mathcal{X}$ and $\mathcal{Y}$ has the same limiting distance after rescaling by the factor $d^{-1/2}$. Then the SVM plane asymptotically perpendicularly bisects any pair of vertices from the two simplices. Denote by X_R the most closest point in $\mathcal{X}$ to the SVM plane, and Y_L that in $\mathcal{Y}$. Then the SVM hyperplane asymptotically perpendicularly bisects the line connecting X_R^* and Y_L^*, where X_R^* and Y_L^* denote the projections on the normal vector of the SVM plane.

By the definition in (1), $\hat{C}^*$ is exactly the middle point of X_R^* and Y_L^* if the two populations are separated. Once the samples overlap, the quantile C^* is a good choice for the cutoff point since it is decided by the distributional structures of the two populations. It is difficult to give mathematically the difference between the RCQ and the SVM rules when the dimensionality and samples are fixed.

However, when $d \to \infty$ the RCQ hyperplane $\mathbf{P}$ convergences to the limiting RCQ plane, $\mathbf{P}_0$ say, it can be shown that the direction of the normal vector $\hat{C}_{XY}$ asymptotically coincides with the direction of the vector $Y_L^* - X_R^*$, and $\hat{C}^*$ is exactly the middle point of the limiting points of X_R^* and Y_L^*, (do a lemma about this in the appendix? ???) after rescaling by the factor $d^{-1/2}$, so that the RCQ plane asymptotically coincides with the SVM plane and each edge of the simplices are asymptotically parallel to $\mathbf{P}_0$. That is, asymptotically for any fixed n_1 and n_2, the RCQ and the SVM rules choose the same separating hyperplane which asymptotically coincides with the DWD plane only when $n_1 = n_2$ (see Hall and Marron 2003).

However, as shown in the next section the errors among DWD, RCQ and SVM rules are totally different. In addition, the DWD and the SVM rules does not cope well with differences of the sampling distributions, especially when the shapes are the same and the variances are different for the two populations; while the RCQ rule robustly estimates the normal vector and automatically adapts to the structures of the sampling distributions. This will be reflected in the simulations in Section 4.

3.2 Error comparison among RCQ, SVM and DWD

The argument employed here for error comparison is basically from Hall and Marron (2003). Let r_i and s_i respectively denote the perpendicular distances from X_i and Y_i to the RCQ hyperplane $\mathbf{P}$ after rescaling by $d^{-1/2}$, similarly let r and s be the distances to $\mathbf{P}$ from the centroids $\hat{C}_X$ and $\hat{C}_Y$, respectively. Then $r = n_1^{-1}\sum_{i=1}^{n_1} r_i$ and $s = n_2^{-1}\sum_{i=1}^{n_2} s_i$. Write $r_i = r_i^0(\mathbf{P}) + \xi_i(\mathbf{P})$ and $s_i = s_i^0(\mathbf{P}) + \eta_i(\mathbf{P})$, where $r_i^0(\mathbf{P})$ and $s_i^0(\mathbf{P})$ are the distances to $\mathbf{P}$ from the simplex vertices to which X_i and Y_i convergence, after rescaling by $d^{-1/2}$, as $d \to \infty$; and $\xi_i(\mathbf{P})$ and $\eta_i(\mathbf{P})$ are stochastic perturbations. Let the finite vectors v and v^0 be parameters for intercepts and slopes of the hyperplane $\mathbf{P}$ and $\mathbf{P}_0$, respectively, where $v = v^0 + d^{-1/2}w$, then following Hall and Marron (2003),

$$r_i^0(\mathbf{P}) = r_i^0(\mathbf{P}_0) + d^{-1/2} w^T \dot{r}_i^0 + o_p(d^{-1/2})$$

and

$$s_i^0(\mathbf{P}) = s_i^0(\mathbf{P}_0) + d^{-1/2} w^T \dot{s}_i^0 + o_p(d^{-1/2}),$$

where $\dot{r}_i^0$ and $\dot{s}_j^0$ denote the vectors of derivatives of $r_i^0(\mathbf{P})$ and $s_j^0(\mathbf{P})$ with respect to v, evaluated at v^0. Then

$$r^0(\mathbf{P}) = r^0(\mathbf{P}_0) + d^{-1/2} w^T \dot{r}^0 + o_p(d^{-1/2})$$

and

$$s^0(\mathbf{P}) = s^0(\mathbf{P}_0) + d^{-1/2} w^T \dot{s}^0 + o_p(d^{-1/2}),$$

where $\dot{r}^0$ and $\dot{s}^0$ denote the averages of $\dot{r}_i^0$ and $\dot{s}_j^0$, respectively. Again following Hall and Marron (2003), $\xi_i(\mathbf{P})$ and $\eta_i(\mathbf{P})$ can be written as $\xi_i(\mathbf{P}) = d^{-1/2}\xi_i^0 + o_p(d^{-1/2})$ and $\eta_j(\mathbf{P}) = d^{-1/2}\eta_j^0 + o_p(d^{-1/2})$, where ξ_i^0 and η_j^0 are independent and normally distributed with mean zero and variance, σ_X^2 and σ_Y^2 say respectively. It follows that

$$r = r_0 + d^{-1/2}(\xi^0 + w^T \dot{r}^0) + o_p(d^{-1/2}),$$

and

$$s = s_0 + d^{-1/2}(\eta^0 + w^T \dot{s}^0) + o_p(d^{-1/2}),$$

where ξ^0, η^0, r^0 and s^0 are respectively the averages of $\xi_i(\mathbf{P})$, $\eta_j(\mathbf{P})$, $r_i^0(\mathbf{P})$ and $s_i^0(\mathbf{P}_0)$. Since r_i^0 and s_i^0 are all the same in the limit, then by the definition of RCQ plane

$$\xi^0 + w^T \dot{r}^0 = \eta^0 + w^T \dot{s}^0 + o_p(1).$$

Then for any fixed n_1, n_2, $w^T(\dot{r}^0 - \dot{s}^0) = \eta^0 - \xi^0$ asymptotically holds. Recall that ξ_i^0 and η_j^0 are independent and normally distributed with mean zero, then when $d \to \infty$

$$w^T(n_1^{-1}\sum_{i=1}^{n_1} \dot{r}_i^0 - n_2^{-1}\sum_{j=1}^{n_2} \dot{s}_j^0) = n_2^{-1}\sum_{j=1}^{n_2} \eta_j^0 - n_1^{-1}\sum_{i=1}^{n_1} \xi_i^0 \approx 0, \tag{12}$$

and the RCQ plane is, up to $o_p(d^{-1/2})$ perturbations of v^0, the plane $\mathbf{P}_0$ after $d^{-1/2}w$ has been added to v^0. Since w in (12) is approximately orthogonal to the vector $n_1^{-1}\sum_{i=1}^{n_1} \dot{r}_i^0 - n_2^{-1}\sum_{j=1}^{n_2} \dot{s}_j^0$ which is a difference between two averages rather than an extremum as for SVM (see Hall and Marron, 2003), the mean square of w is generally smaller than that for SVM, which holds even for unequal n_1 and n_2 while for the DWD this is true only when $n_1 = n_2$. This suggests a reason why the RCQ rule is effective more often than its alternatives in the simulations below.

4 Simulations

This section reports the results of a simulation study comparing the simple centroid, SVM, DWD and RCQ methods for HDLSS classification. The data are essentially iid standard normal, with means $+3/\left(2\sqrt{d}\right)$ for Class $\mathcal{X}$ and means $-3/\left(2\sqrt{d}\right)$ for Class $\mathcal{Y}$. We focus on these important variations in HDLSS setting:

1. Sample Sizes: same ($n_1 = n_2 = 25$) or different ($n_1 = 25,\ n_2 = 50$).

2. Class Variances: same ($\sigma^2 = \tau^2 = (1)^2$) or different ($\sigma^2 = (1)^2,\ \tau^2 = (4)^2$).

3. Population Shape: Spherical Gaussian or with heteroscedasticity where the first half of the entries magnified by a factor of 4, and the second half are shrunken by a factor of 1/4.

4. Population Shape: Standard Gaussian, or with 10% outliers of $10\sqrt{d}$ in the first entry only.

   In the spirit of the mathematical results above, we fixed n_1 and n_2 as indicated in (1) above, and worked with a range of dimensions, $d = 10,\ 25,\ 100,\ 400,\ 1600$.

We computed the classification error rates, for all $2x2x2x2 = 16$ setting above, for all 5 values of d. The full results are too voluminous to report here, so we present summaries, and a detailed look at some of the most interesting cases. RCQ was the best, or among the best in a majority of the cases considered here, suggesting robust performance across a wide variety of cases.

Figure 1 shows a summary of the misclassification rates for a setting where RCQ was generally much better than the other methods. This was for $n_1 = n_2 = 25$, different Class variances ($\sigma^2 = (1)^2,\ \tau^2 = (4)^2$), heteroscedastic population shape, and no outliers.

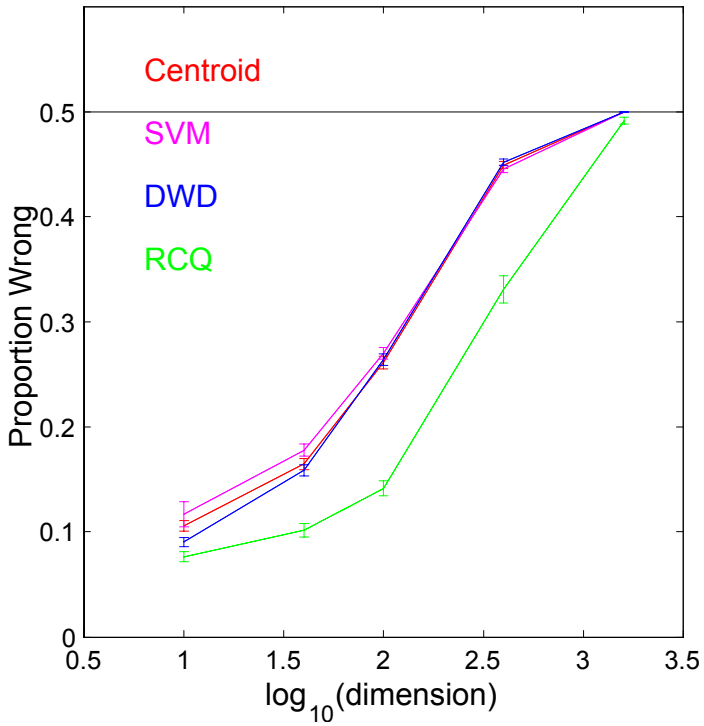

FIGURE 1: *Comparison of Centroid, SVM, DWD, RCQ, for same n, different variance, same shape, no outliers.*

Figure 1 shows that except for the extreme dimensions, $d = 10$ and $d = 1600$, RCQ is substantially better than the other methods. The convergence of all methods is expected for very large d from the asymptotic theory described in Section 3.1.

Figures 2 and 3 give an indication of the reason behind the superior performance on RCQ. They each show the projections of the two classes onto the direction vectors, for RCQ in Figure 2, and for DWD in Figure 3. In each case 4 simulated realizations are shown, to give a simple visual impression of the variation across simulated data sets.

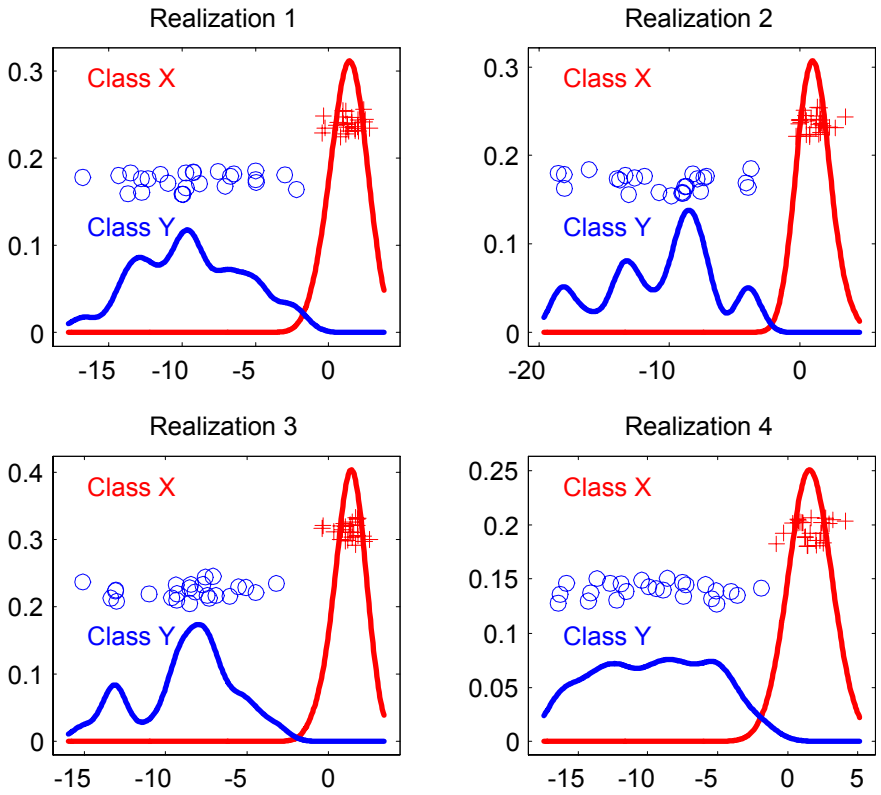

FIGURE 2: *Projection of the training data (4 realizations) onto the direction vector determined by RCQ, in the case of Figure 1.*

Figure 2, showing the RCQ direction, correctly shows the large difference between in variance between the two classes. This is expected from the method that was used to generate the data. This also clearly shows why it is not enough to use the midpoint between the centroids as the cutoff $\hat{C}$, in the algorithm described in Section 2. Instead one should use a cutoff, that correctly balances the spread of the two classes, which was the motivation for $\hat{C}^*$ defined in (1).

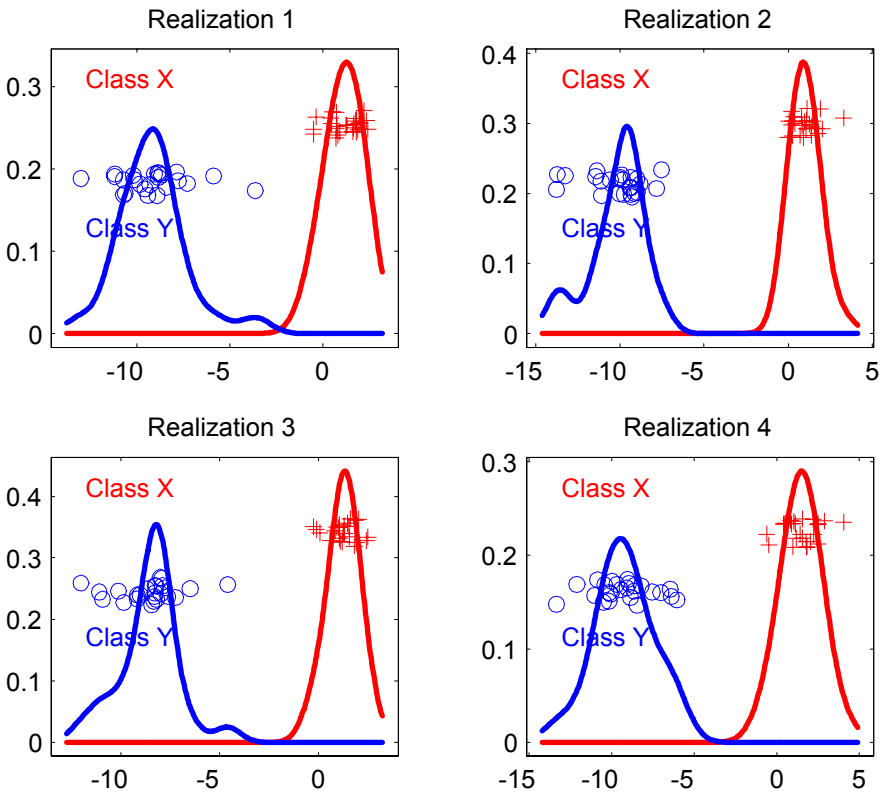

FIGURE 3: *Projection of the training data (4 realizations) onto the direction vector chosen by DWD, in the case of Figure 1.*

Figure 3 shows the corresponding graphic for the DWD direction. Note that in all cases, the classes are "better separated" than in Figure 2, which is not surprising, since DWD attempts to "maximize the separation between the classes". However, note that this is done at some cost in terms of the spread within each class (Class $\mathcal{Y}$ no longer appears 4 times as spread as Class $\mathcal{X}$), which is what gives the superior performance of RCQ, as shown in Figure 1. The analog of Figure 3 for SVM is similarly poor for DWD, where again the respective spread of the classes disappears, because SVM again attempts to maximize separation of the classes. The analog of Figure 2 for the simple centroid method is quite similar to Figure 2 (in particular showing the relative class spreads correctly), but this method is inferior, because it takes the midpoint of the centroids as the cutoff $\hat{C}$, which shows the value of the quantile adjusted version $\hat{C}^*$ defined in (1).

As noted in the introduction, each method had some situations where it was best. Revealing insights come from understanding these.

The simple Centroid was best when the variance was the same ($\sigma^2 = \tau^2 = (1)^2$), the population shapes were homoscedastic, and there were no outliers. This makes sense because in these Gaussian settings, the centroid method is Bayes risk optimal. Otherwise, outliers have a significant impact on the sample mean (the simple centroids), or the centroid midpoint is quite ineffective.

SVM was the best of the methods considered here when the variance was the same ($\sigma^2 = \tau^2 = (1)^2$), but the distribution shapes were heteroscedastic, and outliers were present. SVM was the worst of the four methods in the homoscedastic case with no outliers. This

fits with the very "nonparametric" approach taken by this method.

As shown in Figure 4, DWD was the best for the heteroscedastic case, in the presence of outliers.

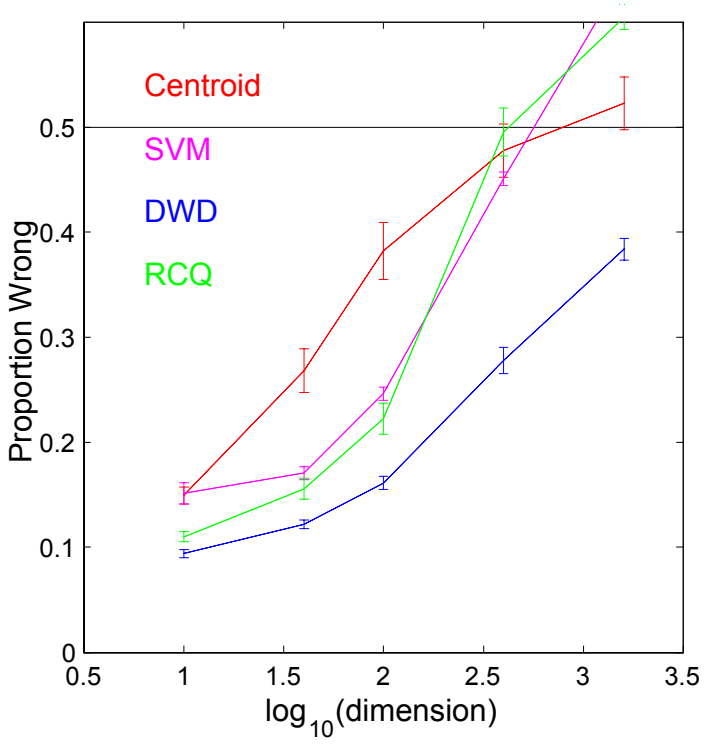

FIGURE 4: *Comparison of Centroid, SVM, DWD, RCQ, for different n, different variance, same shape, outliers.*

It is revealing to study the reason behind the good performance of DWD in this case, using the same graphical device as in Figures 2 and 3. Figure 5 shows the projections of the two classes onto the RCQ directions, again for 4 realizations.

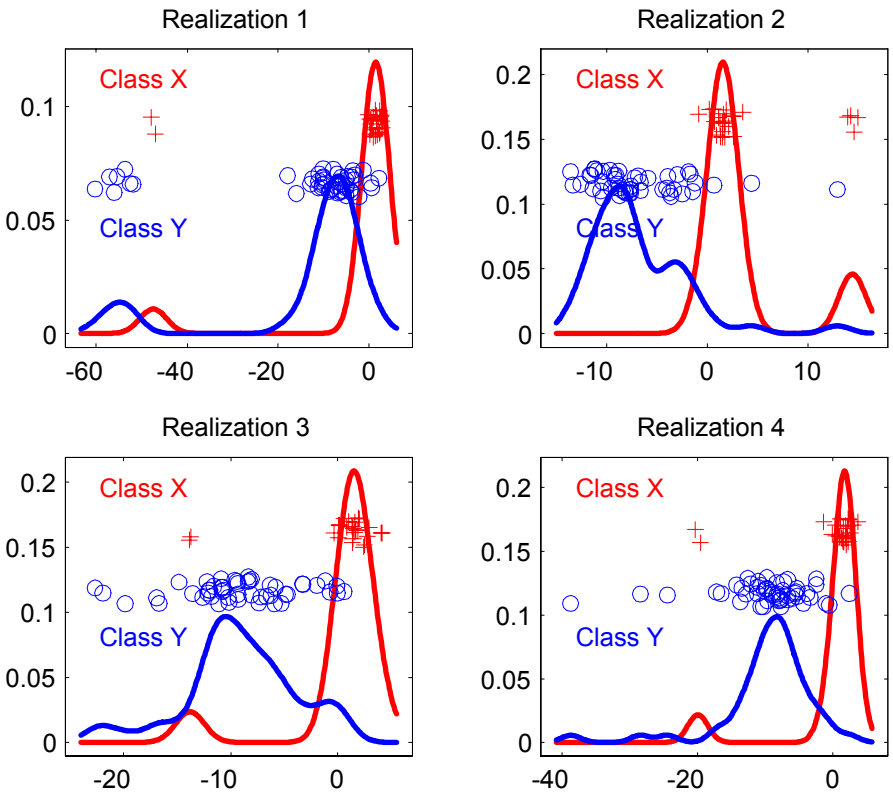

FIGURE 5: *Projection of the training data (4 realizations) onto the direction vector determined by RCQ, in the case of Figure 4.*

Note that this time the 4 projections reveal rather different directions, which shows the RCQ is rather unsteady in this case, suggesting that this direction is not so useful for discrimination (a stark contrast to Figure 2).

Figure 6 shows the corresponding projections onto the DWD directions.

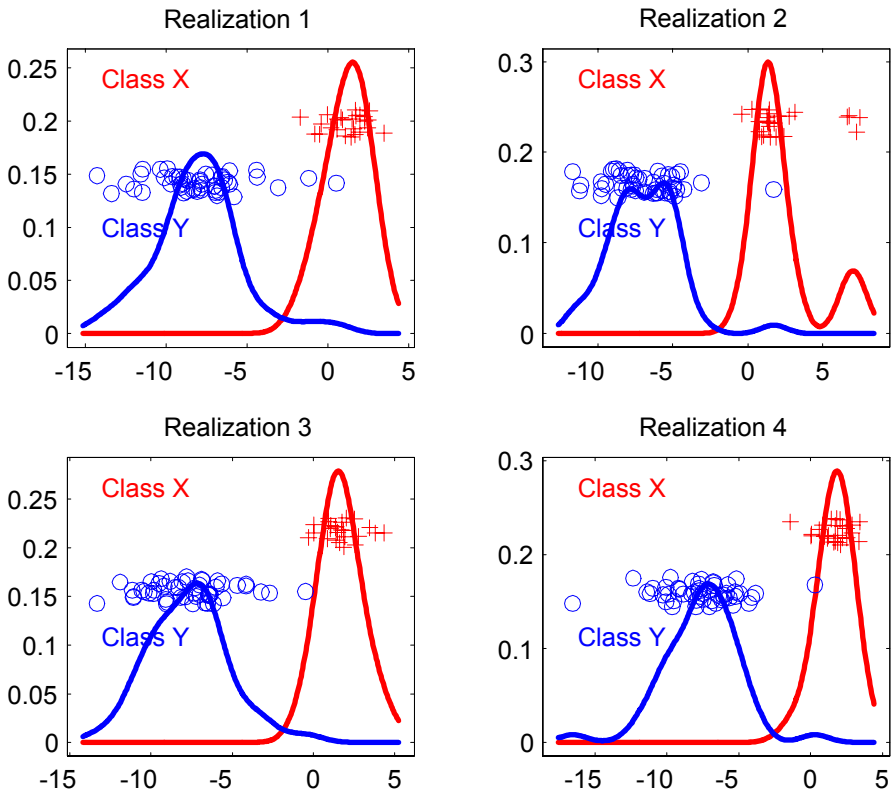

FIGURE 6: *Projection of the training data (4 realizations) onto the direction vector chosen by DWD, in the case of Figure 4.*

Figure 6 shows that for these data, DWD finds a much more useful discrimination direction that RCQ, as shown in Figure 5. This time trying to optimize the separation between the classes is much more useful for classification than the direction chosen the class centroids.

In summary over all cases, RCQ was generally better for $n_1 = n_2$, and also for $\sigma^2 = \tau^2$, with heteroscedastic shape and no outliers, because in this case the robust centroid direction is very good, and the quantile adjusted cutoff $\hat{C}^*$ gave an added advantage. RCQ had the most difficulty for $\sigma^2 = \tau^2$, with homoscedastic shape and outliers, because in that case the outliers had the most influence on the quantiles, yet there was no advantage to the quantile cutoff $\hat{C}^*$.

A variation that we tried, without dramatic success, was to replace the empirical distribution functions by class size weighted versions in (3), i.e. using the cutoff

$$\hat{C}^{**} \equiv \text{median}\{c : \; n_1 F_X^*(c) = n_2 \, (1 - G_Y^*(c))\}, \tag{13}$$

This point could be investigated more deeply, in particular by trying values of n_1 and n_2 which are more different than those considered here.

Another issue that we suspect could be important is the variation in the classification error rates (we only studied the means across our simulations).

References

[1] Duda, R. O., Hart, P.E. and Stork, D. G. (2000) Pattern Classification (2nd Edition), Wiley-Interscience, New York.

[2] Gower,J. C. (1974). The mediancentre. *Applied Statistics*, 23, 466-470.

[3] Haldane, J. B. S. (1948) Note on the median of a multivariate distribution. *Biometrika*, 35, 414-415.

[4] Hall, P. and Marron, J. S. (2003) Geometric Representation of High Dimension Low Sample Size Data, manuscript under preparation.

[5] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust statistics: the approach based on influence functions*, Wiley: New York.

[6] Huber, P. J. (1981) *Robust Statistics*, Wiley: New York.

[7] Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999) Robust Principal Component Analysis for Functional Data, *Test*, 8, 1-73.

[8] Marron, J. S. and Todd, M. J. (2002) Distance Weighted Discrimination, submitted for publication, internet available at: http://www.optimization-online.org/DB_HTML/2002/07/513.html.

[9] Milasevic, P. and Ducharme, G. R. (1987) Uniqueness of the spatial median, *Annals of Statistics*, 15, 1332-1333.

[10] Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust regression and outlier detection*, Wiley: New York.

[11] Staudte, R. G. and Sheather, S. J. (1990) *Robust estimation and testing*, Wiley: New York

[12] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002a) Diagnosis of multiple cancer types by shrunken centroids of gene expression", *Proceedings of the National Academy of the Sciences*, USA, 99, 6567-6572.

[13] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002a) Class prediction by nearest shrunken centroids, with applications to DNA microarrays, internet available at: http://www-stat.stanford.edu/~tibs/ftp/ncshrink2.pdf.

[14] Vapnik, V. N. (1982) *Estimation of dependences based on empirical data*, Springer Verlag, Berlin (Russian version, 1979).

[15] Vapnik, V. N. (1995) *The nature of statistical learning theory*, Springer Verlag, Berlin.