Geometric Representation of High Dimension Low Sample Size Data

Peter Hall
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200
Australia

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

Amnon Neeman
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200
Australia

November 7, 2004

Abstract

High Dimension, Low Sample Size data are emerging in a number of areas of science. In this paper a common structure underlying many such data sets is found, using a non-standard type of asymptotics: the dimension tends to infinity while the sample size is fixed. Our analysis shows a tendency for the data to lie deterministically at the vertices of a regular simplex. Essentially all of the randomness in the data appears only as a random rotation of this simplex. This geometric representation is used to obtain several new statistical insights.

1 Introduction

High Dimension, Low Sample Size (HDLSS) data are becoming increasingly common, in a number of fields. These include genetic micro-arrays, medical imaging and chemometrics, which we treat briefly in the next three paragraphs.

A currently very active area of data analysis is micro-arrays for measuring gene expression; see for example Eisen and Brown (1999), Alter, et al. (2000), Perou et al. (1999, 2000) and Sørlie et al. (2001). A single measurement yields simultaneous expression levels for thousands to tens of thousands of genes. Because the measurements tend to be very expensive, the sizes of most data sets are in the tens, or maybe low hundreds, and so the dimension, d, of the data vectors is much larger than the sample size, n.

In medical image analysis, there are many research problems which currently need statistical input. These lie in the direction of understanding and analyzing populations of three-dimensional images. A useful approach is to first numerically represent shapes of organs of interest. This is done in a wide variety of ways, including the boundary representations developed by Cootes and Taylor (1993), and the completely different medial representations, well described by Yushkevich, et al. (2001). This results in numerical summaries, in the form of vectors of parameters, with dimensionality usually in the high tens to low hundreds for three-dimensional images. However, such representations are often expensive to compute, mostly because the segmentation step (i.e. finding the boundary of the object) often requires at least some human intervention on a slice by slice basis. Thus sample sizes (i.e. numbers of such representations that are collected) are usually in the low tens, again resulting in HDLSS data.

Various types of spectral measurements are very common in chemometrics, where the spectra are recorded in channels that number well into the thousands; see for example Schoonover, Marx, and Zhang (2003) and Marron, Wendelberger and Kober (2004). As with the above fields, practical considerations limit the number of samples to far fewer than the number of channels, again resulting in $n << d$.

Such HDLSS data present a substantial challenge to many methods for classical statistical analysis. Indeed, the first step in a standard multivariate analysis is often to "sphere the data", through multiplying the data matrix by the root inverse of the covariance matrix. For HDLSS data, however, this inverse does not exist, because the covariance matrix is not of full rank.

As part of the development process of new methodologies, there is a need to validate, assess and compare them. For this purpose it is useful to employ both numerical simulation and mathematical analysis. In this paper we provide a mathematical structure within which asymptotics for $d \to \infty$, with n fixed, gives informative and insightful results. The key idea is to study either the subspace or the hyperplane generated by the data. When the data satisfy some fairly standard distributional conditions, the subspace or hyperplane can be rotated in such a way that the data converge to the vertices of a *deterministic* regular simplex. Thus HDLSS data sets, modulo a random rotation, tend towards the latter elementary geometric representation.

The asymptotics in this paper seem to be the first to seriously treat the HDLSS case of $d \to \infty$, with n fixed. The most common case in the current literature is $n \to \infty$, with d fixed. Some researchers, e.g. Huber (1973) and Portnoy (1984, 1988), have addressed the case of $n \to \infty$, with d also growing, say as some power (generally less than one) of n. Bai and Sarandasa (1996),

Sarandasa and Altan (1998) and Johnstone (2001) have studied asymptotics where $n \to \infty$, and d grows at the same rate. The risk bounds of Tsybakov (2003) have very interesting implications across a wide range of combinations of $n \to \infty$ and $d \to \infty$. Rao (1973) discusses some ideas of Mahalanobis (1936), who considered the relationship of populations as $d \to \infty$. See Rao and Varadarajan (1963) for discussion of these issues in the context of stochastic processes.

For simplicity of presentation, these ideas are first explored in the standard Gaussian case, via some elementary calculations, in Section 2. A more general mathematical treatment follows in Section 3.

This new geometric representation is used to analyze the HDLSS performance of some discrimination rules, including the Support Vector Machine, in Section 4. In addition to giving a mathematical tool for comparison of methods, the new geometric representation also provides an explanation for some previously puzzling phenomena.

2 Standard Gaussian Geometrical Representation

Insight into the high dimensional phenomena which drive the geometric representations developed in this paper comes from some perhaps non-obvious facts about high dimensional standard normal distributions. Let $Z(d) = (Z^{(1)}, \ldots, Z^{(d)})^{\mathrm{T}}$ denote a d dimensional random vector drawn from the normal distribution with zero mean and identity covariance matrix. Because the sum of the squared entries has a chi-squared distribution with d degrees of freedom, which tends towards the Gaussian as $d \to \infty$, a simple delta method calculation shows that the Euclidean distance has the following property:

$$\|Z\| = \left\{ \sum_{k=1}^{d} \left(Z^{(k)} \right)^2 \right\}^{1/2} = d^{1/2} + O_p(1).$$

This provides a sense in which the data lie near the surface of an expanding sphere. The result is readily extended to the case of two independent vectors from the standard normal, $Z_1(d)$ and $Z_2(d)$ say:

$$\|Z_1 - Z_2\| = (2d)^{1/2} + O_p(1) \text{ as } d \to \infty. \tag{1}$$

Thus data points tend to be a deterministic distance apart, in a similar sense. A further useful insight comes from considering the angle, at the origin, between the vectors Z_1 and Z_2. Again a simple delta method calculation, this time for the inverse cosine of the inner product, gives:

$$\operatorname{ang}(Z_1, Z_2) = \tfrac{1}{2}\pi + O_p\left(d^{-1/2}\right), \tag{2}$$

where $\operatorname{ang}(Z_1, Z_2)$ denotes the angle, in measured radians at the origin, between vectors Z_1 and Z_2. Of course, both (1) and (2) hold for a random sample

$Z_1, ..., Z_n$, implying that all pairwise distances in the sample are approximately equal and that all pairwise angles are approximately perpendicular. This is challenging to visualize for $n \geq 4$.

These properties are illustrated in Figure 1, where the case $d = 3$, $n = 3$ is considered. All the rays from the origin to the respective data points, shown as solid blue lines, are of approximately equal length, and the distances between data points (measured along the dashed blue lines) are all about $2^{1/2}$ times as large. The rays from the origin are also nearly orthogonal. It is a matter of personal taste whether to focus attention on the subspace generated by the data (of dimension $n = 3$ in this case), or on the hyperplane generated by the data (of dimension $n-1 = 2$ here). Here only the structure of the data in the hyperplane is explored further. Because all pairwise distances are nearly the same, the data lie essentially at the vertices of an equilateral triangle, which is the "regular 3-hedron", i.e. a 3-simplex. This is the picture that will be most useful to keep in mind during the general analysis in Section 3. (A topologist would generally refer to our 3-simplex as a 2-simplex, notating it using the number of dimensions in which it lives, rather than its number of vertices. However, notation in this paper will be simpler if we index a simplex in terms of its number of vertices, and so we shall follow that course.)

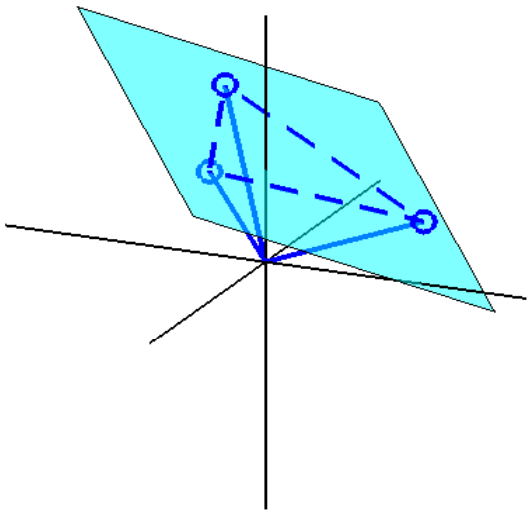

FIGURE 1: *Three point toy example, showing geometric representation, by rotation of the two dimensional hyperplane containing the data, to give a regular n-hedron.*

Another example elucidating these ideas is shown in Figure 2. Each panel shows overlaid scatterplots of 10 samples (shown as different colors) of standard normal random vectors of size $n = 3$, and in $d = 2$, 20, 200 and 20000 dimensions in the respective panels. The 10 samples give an impression of the sampling

variation, as a function of the dimension, which varies for the panels. For each sample, and each dimension, the hyperplane generated by the data (i.e. the plane shown in Figure 1) is found, and the data are projected onto that. Within that hyperplane the data are rotated so that the horizontal coordinates of the bottom two points are centered on 0, to give the scatterplots shown in Figure 2. In view of (1) it is anticipated that these points will lie close to the vertices of the equilateral triangle, with side length $(2d)^{1/2}$ shown with black dashed lines (the regular 3-simplex}, and that this approximation will be better for higher dimensions.

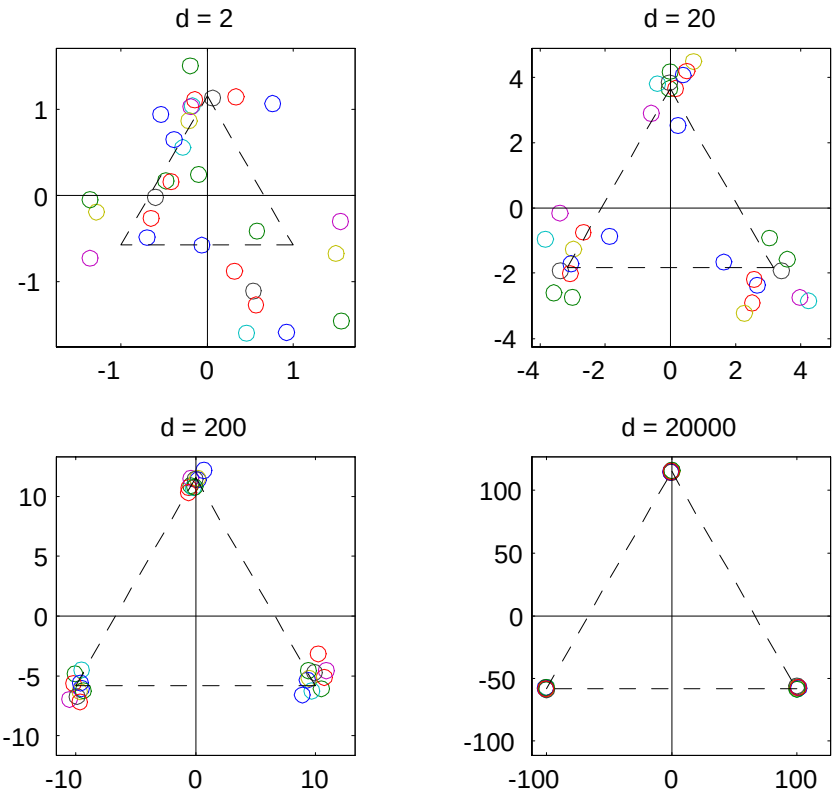

FIGURE 2: *Gaussian Toy Example, illustrating geometric representation, for $n = 3$, and convergence to a 3-simplex with increasing dimension.*

The figure confirms these conjectures. Note that for $d = 2$, the points appear to be quite random, and indeed not all of them are easy to associate with the appropriate vertex of the triangle. However, for $d = 20$ there is reasonable convergence to the vertices, suggesting that the geometric representation is already informative. For $d = 200$ the approximation is quite good, making it clear that the majority of variability goes into the two rotations considered above. As expected, the case $d = 20000$ shows an even more rigid geometric representation.

Andrew Barron remarked that this geometric representation bears a strong similarity to some of the ideas that underly Shannon information theory.

3 General Geometrical Representation

In this section, the geometric representation is made more general. Section 3.1 treats the single sample case. Section 3.2 extends these ideas to two data sets from different distributions, to lay the foundation for using geometric representation ideas for the analysis of discrimination methods.

3.1 Representation of a single sample

Consider a data vector $X(d) = (X^{(1)}, \ldots, X^{(d)})^{\mathrm{T}}$, obtained by truncating an infinite time series which we write as a vector, $X = (X^{(1)}, X^{(2)}, \ldots)^{\mathrm{T}}$. If a law of large numbers applies to the time series, in the sense that $d^{-1} \sum_k \left(X^{(k)}\right)^2 \to a$ in probability, for a constant $a > 0$, then we might fairly say that $X(d)$ lies approximately on the surface of a d-variate sphere, of radius $(ad)^{1/2}$, as $d \to \infty$.

The approximate n-simplex structure, observed in Section 2, will follow from the limiting behavior of distances between pairs of points in a sample, $\mathcal{X}(d) = \{X_1(d), \ldots, X_m(d)\}$, where the data vectors, $X_i(d)$, are taken to be independent and identically distributed as $X(d)$. Assume:

1. The fourth moments of the entries of the data vectors are uniformly bounded.

2. For a constant σ^2,
$$\frac{1}{d} \sum_{k=1}^{d} \mathrm{var}\left(X^{(k)}\right) \to \sigma^2 \,. \tag{3}$$

3. The time series X is ρ-mixing for functions that are dominated by quadratics, as defined in Section 5.1.

Then it follows by a law of large numbers that the distance between $X_i(d)$ and $X_j(d)$, for any $i \neq j$, is approximately equal to $(2\sigma^2 d)^{1/2}$ as $d \to \infty$, in the sense that
$$\frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} \left(X_i^{(k)} - X_j^{(k)}\right)^2 \right\}^{1/2} \to \left(2\sigma^2\right)^{1/2}, \tag{4}$$
where the convergence is in probability. See Section 5.1 for details on (4).

Note that stationarity of the time series X is not required. Instead we only need boundedness of moments, weak independence, and condition (3), which entails stationarity only in a very weak, Césaro-averaged, first-order form. In this sense we are working with a rich class of models for high-dimensional data. Assumption 3 is a simple way of permitting the amount of information available for discrimination to diverge to infinity as d increases. (In conventional

asymptotics, information diverges through increasing sample size.) However, it is also of interest to explore more marginal cases where conditions such as assumption 3 fail; see Section 6.

Application of the result (4) to each pair (i, j) with $1 \le i < j \le m$, and scaling all distances by the factor $d^{-1/2}$, shows that the pairwise differences between points in $\mathcal{X}(d)$ are all asymptotically equal to $(2\sigma^2)^{1/2}$, as $d \to \infty$. Equivalently, if we work with the $(m-1)$-dimensional space into which all m points in $\mathcal{X}(d)$ can be projected without losing their intrinsic relationships to one another, and rescale as before, we conclude that:

$$\begin{array}{l}\text{after rescaling by } d^{-1/2}\text{, the points } X_i(d) \text{ are asymptotically located} \\ \text{at the vertices of an } m\text{-simplex where each edge is of length } (2\sigma^2)^{1/2}.\end{array} \tag{5}$$

Of course, the theory described in (5) involves keeping m fixed as d increases. As noted in Section 2, the m-simplex is an m-polyhedron with all edges of equal length, e.g. for $m = 3$ the equilateral triangle with dashed edges shown in Figure 1.

3.2 Representation of two samples

For the study of classification, the two sample case is also important. Suppose that, in addition to the sample $\mathcal{X}(d)$ where data vectors are distributed as the first d components of the time series X, there is an independent, random sample $\mathcal{Y}(d) = \{Y_1(d), \ldots, Y_n(d)\}$, where each $Y_j(d) = (Y_j^{(1)}, \ldots, Y_j^{(d)})^{\mathrm{T}}$ is distributed as the first d components of a time series $Y = (Y^{(1)}, Y^{(2)}, \ldots)^{\mathrm{T}}$. Straightforward modifications of the assumptions 1–3 in Section 3.1 for the time series Y, together with a new assumption about separation of population means, gives the new conditions:

$$\frac{1}{d} \sum_{k=1}^{d} \mathrm{var} \left(Y^{(k)} \right) \to \tau^2 \,, \quad \frac{1}{d} \sum_{k=1}^{d} \left(EX^{(k)} - EY^{(k)} \right)^2 \to \mu^2 \,, \tag{6}$$

where τ and μ denote finite, positive constants. It follows that the analogue of (4) holds: after rescaling by the factor $d^{-1/2}$, the data $Y_i(d)$ are asymptotically located at vertices of an n-simplex where each edge is of length $2\tau^{1/2}$.

As will shortly be seen, the second part of (6) is especially relevant to accurate classification. If μ in (6) is too small, and in particular if it equals zero, then a classifier of any conventional type (support vector machine, distance weighted discrimination, nearest neighbour, etc) operates asymptotically in a degenerate fashion, without respecting the population, with probability converging to 1 as $d \to \infty$, from which a new datum comes. That is, the classifier assigns the new datum to the same population, regardless of the actual population from which it came. In such instances the classifier is overwhelmed by the stochastic noise that accrues from a very large number of dimensions. The case $\mu = 0$ can arise when there is only a finite number of truly discriminating components.

Since the samples $\mathcal{X}(d)$ and $\mathcal{Y}(d)$ are independent, a weak law of large numbers and property (6) show that the distance between $X_i(d)$ and $Y_j(d)$, divided

by $d^{1/2}$, converges in probability to $(\sigma^2 + \tau^2 + \mu^2)^{1/2}$ as $d \to \infty$:

$$\frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} \left(X_i^{(k)} - Y_j^{(k)} \right)^2 \right\}^{1/2} \to \ell \equiv \left(\sigma^2 + \tau^2 + \mu^2 \right)^{1/2} . \tag{7}$$

See Section 5.1 for details. Thus, after rescaling all distances by the factor $d^{-1/2}$, and writing N for $m + n$, we obtain the following geometric picture of the two samples, $\mathcal{X}(d)$ and $\mathcal{Y}(d)$, for large d and fixed m and n:

> After rescaling each component of d-variate space by the factor $d^{-1/2}$, the N points in $\mathcal{X}(d) \cup \mathcal{Y}(d)$ 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 m of the vertices are the limits of the m points of $\mathcal{X}(d)$, and are the vertices of an m-simplex of edge length $2^{1/2}\sigma$. The other n vertices are the limits of the n points of $\mathcal{Y}(d)$, and are the vertices of an n-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 $\mathcal{X}(d)$ to one deriving from a point in $\mathcal{Y}(d)$, are all of length ℓ. (8)

The results here hold as $d \to \infty$, for fixed m and n. An N-polyhedron is a figure in $(N-1)$-dimensional space that has just N vertices and has all its faces given by hyperplanes in $(N-1)$-variate space. The particular one discussed at (8) 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.

Note that if $\sigma = \tau$ and $\mu = 0$ (e.g. if the time series X and Y have the same distribution) then the N-polyhedron discussed at (8) is exactly an N-simplex, with all edge lengths $\left(2\sigma^2\right)^{1/2}$.

In the general case, the N-polyhedron of the two sample geometric representation can be constructed by rescaling an N-simplex, as follows. An N-simplex has m of its vertices arranged as those of an m-simplex in $(m-1)$-variate space, and the other n vertices arranged in an n-simplex in $(n-1)$-variate space. Alter the scales of these two simplices so that their respective edge lengths are $2^{1/2}\sigma$ and $2^{1/2}\tau$; each is still a simplex in its own right. Then alter the lengths of the other edges, of which there are

$$\tfrac{1}{2}\, N\,(N-1) - \tfrac{1}{2}\, m\,(m-1) - \tfrac{1}{2}\, n\,(n-1) = m\, n\,,$$

so that they all equal ℓ.

Examples for small values of m and n are readily visualized, as discussed in the next paragraph. We shall use the term "tetrahedron" for the non-regular version of that figure, in which edge lengths are not necessarily equal. In the following paragraph we shall write simply $\mathcal{X}$ and $\mathcal{Y}$ for $\mathcal{X}(d)$ and $\mathcal{Y}(d)$, respectively.

When $m = 2$ and $n = 1$ the N-polyhedron is a triangle, with one of its edges being of length $2^{1/2}\sigma$ and the corresponding two vertices representing the points

in $\mathcal{X}$, the other two edges being of length ℓ, and the third vertex representing the single point in $\mathcal{Y}$. When $m = 3$ and $n = 1$ the N-polyhedron is the surface of a tetrahedron, with the three vertices in its base representing the points in $\mathcal{X}$ and forming an equilateral triangle of side length $2^{1/2}\sigma$, and the vertex at the apex representing the point in $\mathcal{Y}$ and being distant ℓ from each of the vertices in the base. When $m = n = 2$ the N-polyhedron is again the surface of a tetrahedron, as follows. Let two of the vertices in the base of the tetrahedron correspond to the two points in $\mathcal{X}$, and let the other vertex in the base, and the vertex at the apex of the tetrahedron, correspond to the two points in $\mathcal{Y}$. Let the edge joining the two $\mathcal{X}$-points be of length $2^{1/2}\sigma$, let the edge joining the other two points be of length $2^{1/2}\tau$, and let the other four edges all be of length ℓ.

This interpretation converts an intrinsically complex, highly stochastic, high-dimensional data configuration into a highly symmetric, virtually deterministic, low-dimensional one. As noted in Section 2, almost all of the stochastic variability in the data goes into random rotation, although some goes into small perturbations of vertices that disappears as $d \to \infty$. As d increases the orientation of the N-polyhedron constantly changes and does not converge in probability. Thus, as $d \to \infty$ the polyhedron is constantly, randomly spinning in a space of ever increasing dimension. Furthermore, the polyhedron's location also varies with d (unless the means are zero, as assumed in Section 2).

4 Analysis of Discrimination Methods

In this section, the geometric representation ideas of Section 3.2 form the basis of a mathematical analysis of observed behavior of discrimination methods. In particular, in the simulation study of Marron and Todd (2002), it was observed that at very high dimensions, the considered techniques all had similar error rates, across a wide array of simulation settings. A basic version of the popular Support Vector Machine (SVM), and the more recently developed method of Distance Weighted Discrimination (DWD) are treated in Section 4.1. Related ideas for other discrimination rules are discussed in Section 4.2. Some of the theoretically predicted effects are more deeply investigated in a small simulation study in Section 4.3.

4.1 Support vector machine and distance weighted discrimination

Several methods for classification operate by dividing the sample union, $\mathcal{X}(d) \cup \mathcal{Y}(d)$, into two classes by a hyperplane, and classifying a new datum as coming from the X- or Y-population according as it lies on one side or the other of the hyperplane. (Here and below, unless otherwise specified, a hyperplane will be $d-1$ dimensional.) When $d \geq N$, and no k data points lie in a $k-2$ dimensional hyperplane (which happens with probability one for data from continuous probability densities), it is always possible to find a hyperplane that has $\mathcal{X}(d)$ entirely on one side and $\mathcal{Y}(d)$ entirely on the other. Attention is restricted to

this "separable" case, and we will study how the different classification methods vary in terms of the hyperplane that they select.

The support vector machine (SVM) method (see e.g. Vapnik, 1982, 1995; Burges, 1998; Christianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2001) has been implemented and studied in a wide variety of different forms. Here we consider only the simplest *basic* version, which chooses the hyperplane that perpendicularly bisects the line segment between the two closest points in the convex hulls of the respective datasets. Note that these points do not have to be data values. In the asymptotic geometric representation described at (8), these convex hulls are precisely the m- and n-simplices, the vertices of which represent the limits, as $d \to \infty$, of the datasets $\mathcal{X}(d)$ and $\mathcal{Y}(d)$, respectively. (Here and below, in a slight abuse of notation, we refer to the limiting simplices of the samples $\mathcal{X}(d)$ and $\mathcal{Y}(d)$ simply as the m-simplex and the n-simplex, respectively.)

It is thus clear that the projection of the basic SVM hyperplane, into the $(N-1)$-dimensional hyperplane generated by the data, where all the data in $\mathcal{X}(d) \cup \mathcal{Y}(d)$ can be considered to lie, is given asymptotically by the unique $(N-2)$-dimensional hyperplane that bisects each of the edges of length ℓ in the N-polyhedron. To illustrate this point, recall from Section 3.2 that when $m = 2$ and $n = 1$ the N-polyhedron is an isosceles triangle, with its base having length $(2\sigma^2)^{1/2}$ and corresponding to the 2-simplex representing the sample $\mathcal{X}(d)$. In this case the projection of the basic SVM hyperplane into the plane of the 3-polyhedron is, in the limit as $d \to \infty$, the straight line that bisects the triangle's two equal sides of length ℓ.

Now add a new random point to d-variate space; it should be independent of the data in $\mathcal{X}(d) \cup \mathcal{Y}(d)$ and have the distribution of either $X(d)$ or $Y(d)$. We claim that:

Theorem 1. *Assume $\sigma^2/m \geq \tau^2/n$; if need be, interchange X and Y to achieve this. If $\mu^2 > (\sigma^2/m) - (\tau^2/n)$, then the probability that a new datum from either the X or the Y population is correctly classified by the basic SVM hyperplane, converges to 1 as $d \to \infty$. If $\mu^2 < (\sigma^2/m) - (\tau^2/n)$, then with probability converging to 1 as $d \to \infty$, a new datum from either population will be classified by the basic SVM hyperplane as belonging to the Y population.*

The proof follows directly from the geometric representation developed in Section 3.2, and is given in Section 5.2.

It follows that for any $\mu \neq 0$, the basic SVM hyperplane gives asymptotically correct classification of new X values whenever m is sufficiently large, for any given value of n; and asymptotically correct classification of new Y values whenever n is sufficiently large, for any given value of m.

Another interesting consequence of Theorem 1 is that if the X and Y populations have the same average variances, i.e. if $\sigma^2 = \tau^2$; and if $\mu^2/\sigma^2 < \left|m^{-1} - n^{-1}\right|$; then the basic SVM classifier ensures asymptotically perfect classification for the population with the larger sample, and asymptotically completely incorrect classification for the population with the smaller sample.

The case of Marron and Todd's (2002) distance weighted discrimination (DWD) approach differs in important respects, at least when the sample sizes m and n are unequal. When $\mathcal{X}(d)$ and $\mathcal{Y}(d)$ are separable as discussed at the beginning of this section, a general version of the DWD hyperplane is defined by minimizing the sum, S_p say, of the pth powers of the inverses of perpendicular distances from a candidate for the hyperplane to points in $\mathcal{X}(d)$ and $\mathcal{Y}(d)$, where $p > 0$ is fixed.

Let us analyse quickly the properties of the DWD hyperplane. Let C_X be the centroid of the simplex $\mathcal{X}(d)$, C_Y the centroid of the simplex $\mathcal{Y}(d)$. It is easy to see that the line joining C_X and C_Y is orthogonal to the linear subspaces generated by the simplices. From this it easily follows that the DWD hyperplane must be orthogonal to the line joining the centroids. Let P be any point on the the interval C_XC_Y. We want to see when it lies on the DWD hyperplane. Relative relationships are diagrammed in Figure 3.

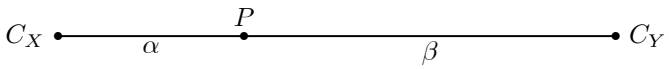

FIGURE 3: *Relative relationships of simplex centroids C_X, C_Y and the candidate DWD cutoff point P.*

Because the simplex $\mathcal{X}(d)$ is orthogonal to C_XC_Y, all the vertices in the simplex are distance α from the hyperplane passing through P, orthogonal to C_XC_Y. Similarly all the points of the simplex $\mathcal{Y}(d)$ are distance β from the hyperplane. The DWD hyperplane minimizes

$$\frac{m}{\alpha^p} + \frac{n}{\beta^p}$$

subject to the constraint that $\alpha+\beta$ is constant. It is an easy exercise in calculus to see that the minimum satisfies the identity

$$\frac{\alpha}{\beta} = \Big(\frac{m}{n}\Big)^{1/(p+1)}. \tag{9}$$

This tells us the location of the DWD hyperplane. It is the hyperplane orthogonal to the line C_XC_Y, passing through the point P which satisfies (9). In section 5.3 we will see how to compute on which side of the hyperplane a new datum point lies. Here we note that $\alpha = \beta$ if and only if $m = n$. In this case, the basic SVM hyperplane and the DWD hyperplane coincide. The larger m/n, the closer the point P will be to C_Y. As $m/n \to \infty$, the DWD hyperplane moves ever closer to the simplex whose vertices represent the smaller of the two samples.

Therefore, Theorem 1 applies without change to the DWD algorithm, provided the two sample sizes are equal. In the contrary case the limit, as $d \to \infty$, of the probability that a new datum is classified as being from the same population as the larger sample, increases with the larger sample size for a fixed

value of the smaller sample size. This anticipates the often-assumed property that the larger sample comes from a population with higher prior probability. In the general case:

Theorem 2. *Assume $\sigma^2/m^{(p+2)/(p+1)} \geq \tau^2/n^{(p+2)/(p+1)}$; if need be, interchange X and Y to achieve this. If $\mu^2 > (n/m)^{1/(p+1)}\,(\sigma^2/m) - (\tau^2/n)$, then the probability that a new datum from either the X or the Y population is correctly classified by the DWD hyperplane, converges to 1 as $d \to \infty$. If $\mu^2 < (n/m)^{1/(p+1)}\,(\sigma^2/m) - (\tau^2/n)$, then with probability converging to 1 as $d \to \infty$, a new datum from either population will be classified by the DWD hyperplane as belonging to the Y population.*

See section 5.3.

Note that, as $p \to \infty$, Theorems 1 and 2 become identical. More generally, the rules which determine success or failure of classification, using basic SVM or DWD, are similar when p is large. In this sense, basic SVM can be viewed as a limiting case of DWD; basic SVM may be regarded as a form of DWD, using a very large value of the exponent that is applied to distance from the space-splitting hyperplane.

Recall from Section 3 that our geometric representations are based on large-d laws of large numbers. The small, stochastic perturbations in those laws are generally asymptotically normally distributed and of size $d^{-1/2}$. An examination of the nature of the perturbations shows that when $m = n$ the DWD hyperplane is less stochastically variable than its basic SVM counterpart, giving rise to the lower error rates for classification. Specifically, stochastic errors in locating the basic SVM hyperplane are, to first order, the result of extrema of small, independent, zero-mean errors in locating simplex vertices. On the other hand, errors in the position of the DWD hyperplane arise from averaging those errors. Since the extrema of independent perturbations are generally larger than the perturbations' average, except in very heavy-tailed cases which are excluded by our moment conditions, then the DWD algorithm produces a less stochastically variable approximation to the common hyperplane to which the basic SVM and DWD hyperplanes converge as $d \to \infty$. This explains the result observed in Figure 5 of Marron and Todd (2002), that for spherical Gaussian data DWD gave somewhat better classification performance than basic SVM.

4.2 Other discrimination rules

Let $C_X(d)$ and $C_Y(d)$ denote the centroids of the datasets $\mathcal{X}(d)$ and $\mathcal{Y}(d)$, respectively. The "centroid rule" or "mean difference rule" classifies a new datum, Z say, as being from the X- or Y-population according as Z is closer to $C_X(d)$ or $C_Y(d)$, respectively. Clearly, $C_X(d)$ and $C_Y(d)$ converge, after rescaling by $d^{-1/2}$ and letting $d \to \infty$, to the centroids of the respective simplices. It follows that the centroid rule discriminator (CRD) enjoys the same properties, described by Theorem 1, as the basic SVM classifier. Indeed, the hyperplane which bisects all the lines (of equal length, ℓ) linking points in the m- and n-simplices, also has the property that it divides space into points which lie nearer to one or other of

the centroids of either simplex. That is, the limit of the basic SVM hyperplane splits space in exactly the same way as the limit of the CRD hyperplane. However, as with DWD, the variation in CRD is driven by averaging the stochastic errors, not by the extrema. This is a new way of understanding the superior performance of CRD over the basic SVM in the example considered in Figure 5 of Marron and Todd (2002). DWD gave essentially the same performance in that case because the sample sizes were equal.

The standard one-nearest-neighbor rule, which classifies Z as coming from the X- or Y-population according as the nearest point in $\mathcal{X}(d) \cup \mathcal{Y}(d)$ is from $\mathcal{X}(d)$ or $\mathcal{Y}(d)$, respectively, has quite different behavior. Instead of Theorem 1 the nearest-neighbor discriminator (NND) satisfies:

Theorem 3. *Assume $\sigma^2 \geq \tau^2$; if need be, interchange X and Y to achieve this. If $\mu^2 > \sigma^2 - \tau^2$, then the probability that a new datum from either the X or the Y population is correctly classified by the NND hyperplane, converges to 1 as $d \to \infty$. If $\mu^2 < \sigma^2 - \tau^2$, then with probability converging to 1 as $d \to \infty$, a new datum from either population will be classified by the NND hyperplane as belonging to the Y population.*

The contrast between Theorems 1 and 3 is marked. For example, taking $m = n$ for simplicity, Theorem 1 asserts that, in the large-d limit, basic SVM misclassifies data from at least one of the populations only when $\mu^2 < \left|\sigma^2 - \tau^2\right| / m$, whereas Theorem 3 asserts that NND leads to misclassification, for data from at least one of the populations, both in this range and when $\left|\sigma^2 - \tau^2\right| / m \leq \mu^2 < \left|\sigma^2 - \tau^2\right|$. This quantifies the inefficiency that might be expected from basing inference on only a single nearest neighbor. Furthermore, without the condition $m = n$, basic SVM has an asymptotic advantage over NND, in the sense of leading to correct classification of data from the X population for a wider range of values of μ, whenever $1 < \tau^2/\sigma^2 < (1 - m^{-1})(1 - n^{-1})^{-1}$; and has this advantage for the Y population if $1 < \sigma^2/\tau^2 < (1 - n^{-1})(1 - m^{-1})^{-1}$.

As noted earlier in this section and in section 4.1, the CRD and DWD (for $m = n$, or for large p) classifiers are equivalent to basic SVM, then the remarks in the previous paragraph remain true if we replace basic SVM by either DWD or CRD. This explains the observation of Marron and Todd, that these methods all gave similar simulation results for very large dimension d (in the case of $m = n$). Furthermore, the four classifiers considered here divide naturally into two groups. The first group contains basic SVM, DWD (for $m = n$ or large p) and CRD, which for large d have similar performance in a wide range of circumstances; and the second group contains just NND, which is generally somewhat inferior to the other two, in terms of the width of the range where it gives correct classification. These issues are illustrated using simulations in Section 4.3.

We have avoided treating "marginal" cases, in particular $\mu^2 = |\sigma^2 m^{-1} - \tau^2 n^{-1}|$ in the setting of Theorem 1 and $\mu^2 = |\sigma^2 - \tau^2|$ in the case of Theorem 3. There the probabilities of misclassification depend on relatively detailed properties of the sampling distribution. Indeed, they are influenced by the

errors in the laws of large numbers which led to properties such as Theorem 1. These errors are generally asymptotically normally distributed, and their joint limiting distributions determine large-d classification probabilities when $\mu^2 = |\sigma^2 m^{-1} - \tau^2 n^{-1}|$ or $\mu^2 = |\sigma^2 - \tau^2|$.

4.3 Simulation illustration

Some of the consequences of the geometric representation ideas developed here are illustrated via simulation in this section.

An interesting, and at the time surprising, observation of the simulation study of Marron and Todd (2002) was that in a variety of simulation settings considered there, for all of basic SVM, DWD and CRD, classification error rates tended to come together for large d. Figure 4 is similar to Figure 5 of Marron and Todd, except that NND has now been added. This shows overall error rates, for the 4 classification methods considered in this paper. Here the training sample sizes were $m = n = 25$, and dimensions $d = 10,\ 40\ , 100,\ 400,\ 1600$ were considered, and the data are standard normal (i.e. multivariate Gaussian with mean 0 and identity covariance), except that the mean of $X_i^{(1)}$, $i = 1, ..., m$ (and of $Y_j^{(1)}$, $j = 1, ..., n$) has been shifted to 2.2, (and -2.2, respectively). Classification error rates were computed based on 100 new data points from each of the two classes, and the means are summarized as the colored curves in Figure 4. Monte Carlo variation, over 1000 repetitions of each experiment, is reflected by the error bars, which are standard normal theory 95% confidence intervals for the true underlying population means.

This simulation setting is not identical to that of this paper, because the first entry of the data vectors have a different mean from the other entries. However, the data space can be simply rotated (through a change of variables) so that the first dimension lies in the direction of the vector whose entries are all 1. Thus this simulation setting is equivalent to the assumptions above, with $\mu = 4.4/d^{1/2}$. In view of the geometrical representation and the calculations in Sections 4.1 and 4.2, it is not surprising that this effectively decreasing value of μ gives error rates that increase in d. Also as expected from the theory, the error rates for basic SVM, DWD and CRD come together for increasing d, although the convergence is perhaps faster than expected. (Recall, from Theorems 1 and 2 and the first paragraph of Section 4.2, that in the case $m = n$ which we are considering here, the classification probabilities for basic SVM, DWD and CRD all converge to 1 as d increases.) Finally, again as predicted, basic SVM lags somewhat behind DWD and CRD (which are not significantly different).

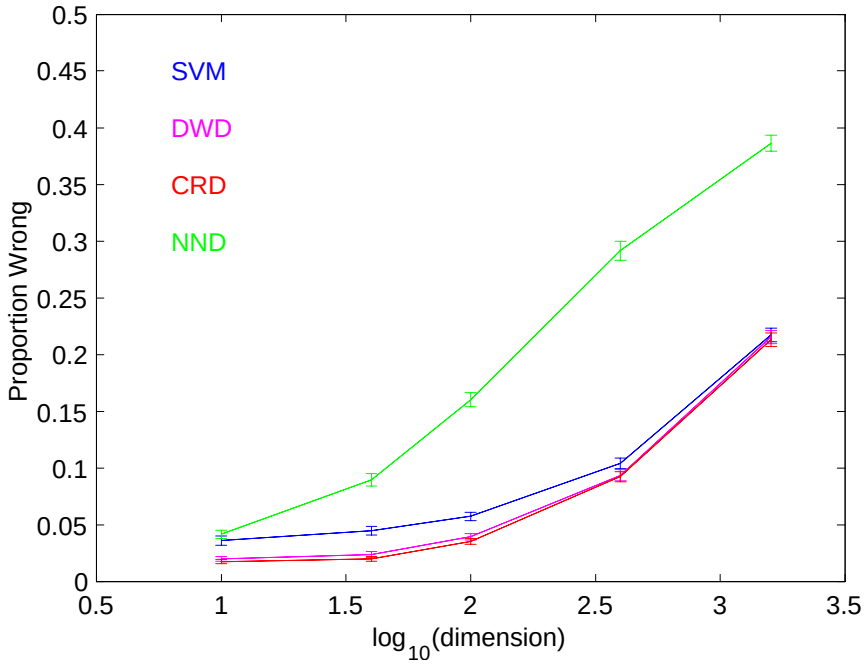

FIGURE 4: *Summarization of simulation results, for Gaussian Data. Shows convergence of most methods for large dimension.*

Simulation performance of the NND rule is also included in Figure 4. As predicted in Section 4.2, NND lags quite substantially behind the other rules in performance (again reflecting the loss in efficiency from using only one nearest neighbor).

The ideas of Theorems 1, 2 and 3 are illustrated in a different way in Figure 5. The simulation setting of Figure 5 is again Gaussian, with training sample sizes $m = n = 16$. This time the parameters are μ as shown on the horizontal axis, $\sigma^2 = 20$ and $\tau^2 = 4$. A range of dimensions, $d = 10,\ 100,\ 1000$, is shown using line type. Classification methods are distinguished using colors and line thickness. Different line thicknesses are used to decrease overplotting effects, e.g. note that for $d = 1000$, basic SVM, DWD and CRD are essentially on top of each other. Again error rates are computed using from 100 new test cases for each class, and averaged over 1000 Monte Carlo repetitions.

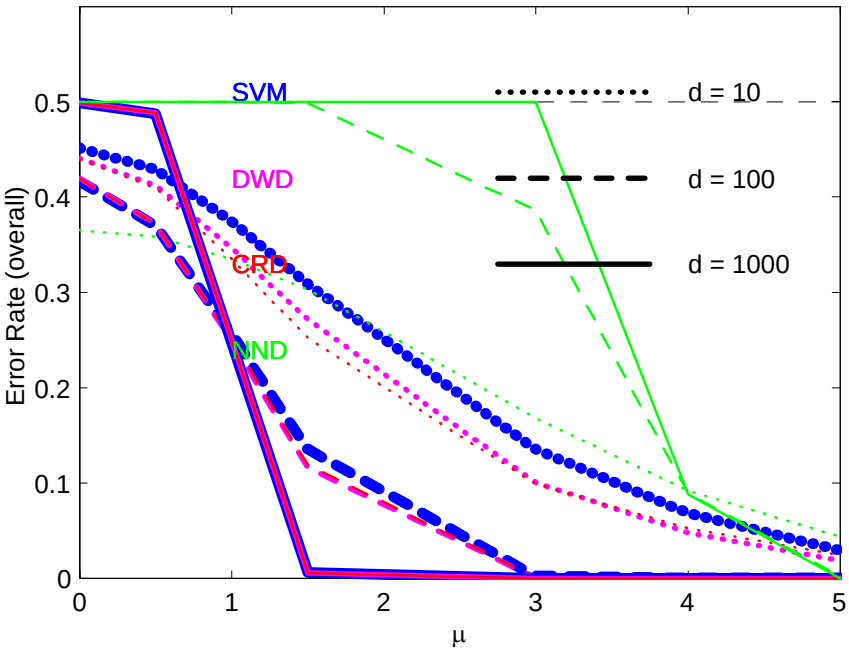

FIGURE 5: *Summary of simulations exploring asymptotic lessons. Shows "change points" at predicted values of* μ.

Figure 5 allows convenient study of classification error rate as a function of μ. Theorem 1 suggests that for "μ large" perfect discrimination (i.e. 0 error rate) is possible for basic SVM, which is reflected by the blue curves coming to 0 on the right side. The convergence is faster for larger dimension d, also as expected. But much more precise information is given in Theorem 1, with in particular a change point at $\mu = \left(\sigma^2/m - \tau^2/n\right)^{1/2} = (20/16 - 4/16)^{1/2} = 1$ expected. To the left of the change point, the theory predicts that the error should be 0.5, because the class $\mathcal{X}(d)$ data will be completely correctly classified, and the class $\mathcal{Y}(d)$ data will all be incorrect. The change point is quite sharp for $d = 1000$ and less so for lower d, as expected, because the geometric representation has not fully taken over for the lower dimensions.

Very similar performance is predicted for DWD by Theorem 2, and seen in Figure 5 as the magenta curve. Performance is virtually identical to basic SVM for $d = 1000$, and again as predicted at the end of Section 4.1, DWD is marginally better for $d = 10,\ 100$.

Recall from Theorem 3 that for NND the change point is quite different, now appearing for $\mu = \left(\sigma^2 - \tau^2\right)^{1/2} = (20 - 4)^{1/2} = 4$ (farther to the right, reflecting the expected inefficiency of 1 Nearest Neighbor Discrimination). This changepoint is also well reflected in Figure 5, as the green curves. Again the asymptotically predicted results are strongest for the highest dimension $d = 1000$.

Our results also suggested that interesting effects should appear for unequal sample sizes n and m. Some simulations in that case are summarized in Figure 6. Specific results are shown for the case of $m = 2$ and $n = 5$. Fairly similar results were obtained for other values of $m = 2$ and $n = 5$.

As in Figure 5, line types represent dimension, $d = 10$, 100, 1000. The results are easiest to interpret when the error rates are broken down in terms of class, so linetype is used to indicate this, with thin lines representing error rates for class $\mathcal{X}(d)$ only, medium lines for class $\mathcal{Y}(d)$ only and the thickest line for the combined error rates. The distributions are again independent Gaussian, with $EX^{(k)} = 0$ and $EY^{(k)} = \mu$, and the variances were taken to be $\sigma^2 = \tau^2 = 1$. Again error rates are displayed as a function of $\mu = \left[\frac{1}{d}\sum_{k=1}^{d}\left(EX^{(k)} - EY^{(k)}\right)^2\right]^{1/2}$.

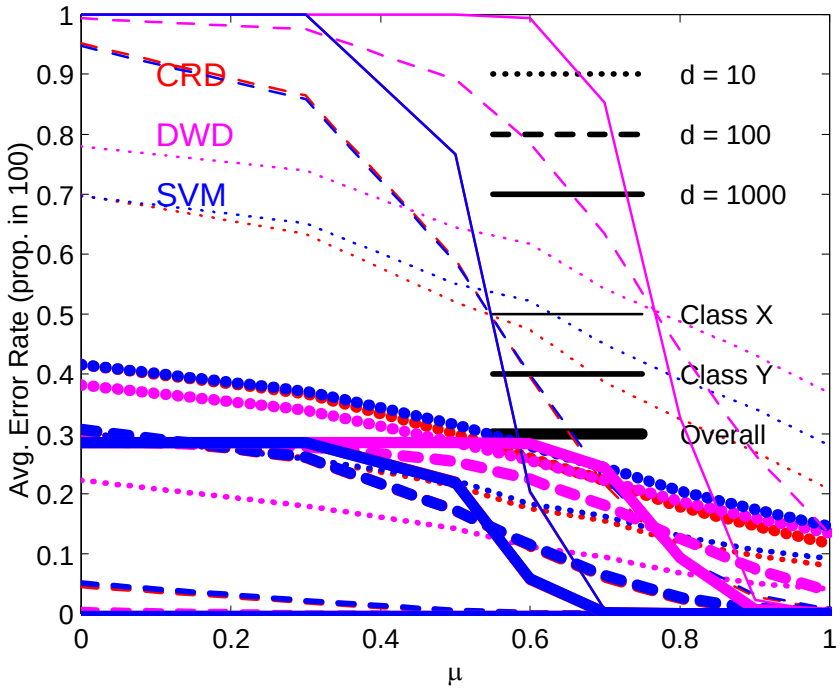

FIGURE 6: *Summary of simulations exploring asymptotic lessons for unequal sample sizes $m = 2$ and $n = 5$. Shows "change points" at predicted values of μ.*

Once again the asymptotically predicted lessons apply. In particular, for small values of μ, the $m = 2$ class $\mathcal{X}(d)$ error rates (indicated by the thin curves) are quite large, and increase to 1 for $d = 1000$ (the thin solid curves). The $n = 5$ class $\mathcal{Y}(d)$ error rates (indicated by medium width lines) are much smaller, and decrease to all 0 for $d = 1000$ (the medium solid curves). The overall rates lie between these, because they are just 2/7 of the class $\mathcal{X}(d)$ rates plus 5/7

of the class $\mathcal{Y}(d)$ rates. For larger values of μ, there is more discrimination information in the data, so all of the rates decrease, with fastest decrease for $d = 1000$ (shown using solid lines).

Also as predicted, for the highest $d = 1000$, CRD and SVM give essentially the same results (i. e. for each line thickness, the solid blue curve is always on top of the corresponding solid red curve), and DWD is substantially worse in terms in terms of both class $\mathcal{X}(d)$ and overall error rates (the thin and thick solid purple curves are higher than the corresponding curves of other colours). For lower dimensions, the results are fairly similar, but DWD seems to give better class $\mathcal{Y}(d)$ performance (shown by the purple medium width lines being below the other two colours), as expected, since DWD errs by using the wrong intercept. DWD even has slightly better overall performance for small values of μ (indicated by the purple thick dotted curve being below the other colours).

The overall error rates (shown as the thick curves), also indicate the predicted performance. For SVM (blue curves), and $\mu > \left(\sigma^2/m - \tau^2/n\right)^{1/2} = (1/2 - 1/5)^{1/2} \approx 0.55$, the error rates tend toward 0 as the dimension d increases. The inferior performance of DWD, predicted by the different threshold of $\mu > \left[(n/m)^{1/(1+p)}\, \sigma^2/m - \tau^2/n\right]^{1/2} = \left[(5/2)^{1/(1+1)}\, 1/2 - 1/5\right]^{1/2} \approx 0.77$, is also clear.

5 Technical Details

This section gives technical details used in the above discussion.

5.1 Laws of Large Numbers

This section gives a concise formulation of the ρ mixing condition, and shows how it can be used to develop the laws of large numbers (4) and (7).

We say that the time series $X = (X^{(1)}, X^{(2)}, \ldots)$ and $Y = (Y^{(1)}, Y^{(2)}, \ldots)$, assumed to be independent of one another and to have uniformly bounded fourth moments, are ρ mixing for functions dominated by quadratics, if, whenever functions f and g of two variables satisfy $|f(u,v)| + |g(u,v)| \le C\, u^2 v^2$ for fixed $C > 0$ and all u, v, we have:

$$\sup_{1 \le k,\ell < \infty,\, |k-\ell| \ge r} \left| \mathrm{Corr}\left\{ f\left(U^{(k)}, V^{(k)}\right),\, g\left(U^{(k)}, V^{(k)}\right) \right\} \right| \le \rho(r)\,,$$

for $(U, V) = (X, X)$, (Y, Y) or (X, Y), where the function ρ satisfies $\rho(r) \to 0$ as $r \to \infty$. See, for example, Kolmogorov and Rozanov (1960).

If the ρ-mixing condition holds, then, by elementary moment calculations,

$$E\left[\sum_{k=1}^{d} \left\{ \left(U_i^{(k)} - V_j^{(k)}\right)^2 - E\left(U_i^{(k)} - V_j^{(k)}\right)^2 \right\}\right]^2 = o\left(d^2\right)$$

as $d \to \infty$, for $(U, V) = (X, X)$, (Y, Y) or (X, Y), where $i \neq j$ if $(U, V) = (X, X)$ or (Y, Y). Therefore, by Chebyshev's inequality,

$$\frac{1}{d} \sum_{k=1}^{d} \left\{ \left(U_i^{(k)} - V_j^{(k)} \right)^2 - E \left(U_i^{(k)} - V_j^{(k)} \right)^2 \right\} \to 0$$

in probability. This result, together with (3) and (6), implies (4) and (7).

5.2 Derivation for basic SVM

This section contains the details leading to Theorem 1.

Let the new datum have the distribution of $X(d)$ and be independent of the data in $\mathcal{X}(d) \cup \mathcal{Y}(d)$. Denote it by $X'(d)$. The asymptotic theory described in sections 3.1 and 3.2 implies that, as $d \to \infty$, the distance of $X'(d)$ from each $X_i(d) \in \mathcal{X}(d)$, rescaled by $d^{-1/2}$, converges in probability to $(2\sigma^2)^{1/2}$; and the rescaled distance of $X'(d)$ from each $Y_j(d) \in \mathcal{Y}(d)$ converges in probability to ℓ.

Recall that we refer to the limiting simplices of the samples $\mathcal{X}(d)$ and $\mathcal{Y}(d)$ as the m-simplex and the n-simplex, respectively. The squared distance from any vertex of the m-simplex to its centroid equals $\sigma^2 (1 - m^{-1})$. To appreciate why, let us temporarily take $\sigma^2 = 1$ and represent the m-simplex in m-variate Euclidean space through its vertices, at the points with coordinates $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$. (This m-variate representation is simpler than an $(m-1)$-variate representation.) Then the centroid of the simplex has coordinates $(m^{-1}, \ldots, m^{-1})$, and so its squared distance from any of the vertices equals $(1 - m^{-1})^2 + (m-1)\, m^{-2} = 1 - m^{-1}$.

Let $Z \in \mathbb{R}^d$ be a point which is distant r from each vertex of the m-simplex. Then Z, any vertex V of the m-simplex, and the centroid of the m-simplex, are the vertices of a right-angled triangle of which the hypotenuse is the line joining Z to V. Therefore, by Pythagoras's theorem, the squared distance from Z to the centroid equals $r^2 - \sigma^2 (1 - m^{-1})$.

The datum $X'(d)$ is correctly classified if and only if it is nearer to the convex hull of the m-simplex than to the hull of the n-simplex. Equivalently, $X'(d)$ is classified as coming from $\mathcal{X}(d)$ if and only if it is nearer to the centroid of the m-simplex than to the centroid of the n-simplex. In view of the result derived in the previous paragraph, the squared distance of $X'(d)$ from the centroid of the m-simplex, and from the centroid of the n-simplex, equal respectively

$$2\sigma^2 - \sigma^2 (1 - m^{-1}) = \sigma^2 (m+1)/m\,, \quad \ell^2 - \tau^2 (1 - n^{-1}) = \mu^2 + \sigma^2 + \tau^2 n^{-1}\,.$$

Hence, $X'(d)$ will be nearer to the n-simplex (and therefore misclassified) if $\sigma^2 (m+1)/m > \mu^2 + \sigma^2 + \tau^2 n^{-1}$, i.e. if $\mu^2 < \sigma^2 m^{-1} - \tau^2 n^{-1}$; and will be nearer to the m-simplex (and so correctly classified) if $\mu^2 > \sigma^2 m^{-1} - \tau^2 n^{-1}$.

So far we have made no assumption regarding which of σ^2/m and τ^2/n is bigger. Now assume $\sigma^2/m > \tau^2/n$. The above tells us when a datum point of type $X'(d)$ will be classified correctly. For a datum point of type $Y'(d)$, the same argument with X and Y interchanged tells us that a datum point of type

Y will be classified correctly if $\mu^2 > \tau^2 n^{-1} - \sigma^2 m^{-1}$. Since the right hand side is negative, this always happens. In other words a datum point of type Y is always classified correctly. Theorem 1 simply assembles together the information about data points of type X and Y.

5.3 Derivation for DWD

In section 5.2 we saw that, given a point Z whose distance from each vertex of the m-simplex $\mathcal{X}(d)$ is r, the squared distance of Z from the centroid of the m-simplex is $r^2 - \sigma^2(1 - m^{-1})$. We can apply this where $Z = Y$ is one of the vertices of the simplex $\mathcal{Y}(d)$. The square of the distance from Y to a point in $\mathcal{X}(d)$ is $\mu^2 + \sigma^2 + \tau^2$, and hence the square distance of Y from the centroid C_X of $\mathcal{X}(d)$ is

$$\mu^2 + \sigma^2 + \tau^2 - \sigma^2(1 - m^{-1}) = \mu^2 + (\sigma^2/m) + \tau^2.$$

Now this is true for every vertex Y in $\mathcal{Y}(d)$. The same analysis now tells us that the square distance of C_X from the centroid C_Y of the simplex $\mathcal{Y}(d)$ is given by

$$\mu^2 + (\sigma^2/m) + \tau^2 - \tau^2(1 - n^{-1}) = \mu^2 + (\sigma^2/m) + (\tau^2/n).$$

Now let $X'(d)$ be a new datum point of type X, independent of $\mathcal{X}(d) \cup \mathcal{Y}(d)$. In section 5.2 we computed the square distances of $X'(d)$ from C_X and C_Y. In other words, in the triangle shown in Figure 6, we know the distances $C_X C_Y$, $X'(d)C_X$ and $X'(d)C_Y$.

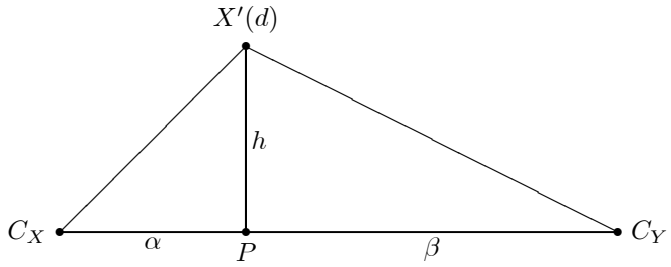

FIGURE 6: *Relative relationships between the new datum point $X'(d)$ and the simplex centroids C_X, C_Y.*

In Figure 6, P is the projection of $X'(d)$ to the line $C_x C_Y$. The distances we have computed tell us

$$\alpha^2 + h^2 = \sigma^2(1 + m^{-1}) \tag{10}$$

$$\beta^2 + h^2 = \mu^2 + \sigma^2 + (\tau^2/n) \tag{11}$$

$$(\alpha + \beta)^2 = \mu^2 + (\sigma^2/m) + (\tau^2/n) \tag{12}$$

Subtracting (11) from (10) we have

$$\alpha^2 - \beta^2 = (\sigma^2/m) - \mu^2 - (\tau^2/n) \tag{13}$$

Adding (12) and (13), and subtracting these two equations, we obtain respectively

$$\alpha(\alpha+\beta) = \sigma^2/m \tag{14}$$

$$\beta(\alpha+\beta) = \mu^2 + (\tau^2/n) \tag{15}$$

from which we conclude that

$$\frac{\alpha}{\beta} = \frac{\sigma^2/m}{\mu^2 + (\tau^2/n)} \tag{16}$$

The point $X'(d)$ will be classified as belonging to X if it lies on the same side of the DWD hyperplane as C_X, that is if

$$\frac{\sigma^2/m}{\mu^2 + (\tau^2/n)} < \Big(\frac{m}{n}\Big)^{1/(p+1)}.$$

It will be classified as belonging to Y if

$$\frac{\sigma^2/m}{\mu^2 + (\tau^2/n)} > \Big(\frac{m}{n}\Big)^{1/(p+1)}.$$

So far our treatment has been general. Now assume $\sigma^2/m^{(p+2)/(p+1)} \geq \tau^2/n^{(p+2)/(p+1)}$. The analysis above tells us when a point $X'(d)$ will be classified correctly. Suppose we have a point $Y'(d)$. By the inequality above

$$\frac{\tau^2/n}{\sigma^2/m} \leq \Big(\frac{n}{m}\Big)^{1/(p+1)}.$$

But then for any positive μ^2 we have

$$\frac{\tau^2/n}{\mu^2 + (\sigma^2/m)} < \frac{\tau^2/n}{\sigma^2/m} \leq \Big(\frac{n}{m}\Big)^{1/(p+1)}.$$

that is, $Y'(d)$ will always be classified as belonging to Y.

Theorem 2 simply combines the information above, on $X'(d)$ and $Y'(d)$.

5.4 Derivation for NND

As in section 5.2, let $X'(d)$ denote a new datum, from the X population, added to the d-variate hyperplane. In the limit as $d \to \infty$, and after the usual normalisation, $X'(d)$ converges to a point whose squared distance from points of the m- and n-simplices equal $2\sigma^2$ and ℓ^2, respectively. Hence, the limit of the probability that $X'(d)$ is correctly classified equals 1 or 0 according as $2\sigma^2 < \ell^2$ or $2\sigma^2 > \ell^2$, respectively. Since $2\sigma^2 < \ell^2$ if and only if $\mu^2 > \sigma^2 - \tau^2$, Theorem 3 follows.

6 Summary and Conclusions

We have shown that, in a model where components of data vectors follow a time series that is stationary in a second-order, Cesàro-averaged sense (see (3) and (6)), the performances of different classifiers for very high dimensions can be represented, quite simply, in terms of the relationships among the average coordinate variances, and the average squared differences of means. This analysis has revealed a variety of different properties of different classifiers. For example, it has been shown that basic SVM and DWD classifiers perform similarly when sample sizes are the same, but not necessarily when sample sizes differ; and that, from some perspectives, basic SVM can be viewed as the limit of DWD as the exponent, p, in the latter increases. It quantifies the belief that, relative to basic SVM and DWD, the NND classifier is swamped by the effects of variability in high-dimensional samples, since (in the case of equal sample sizes) the condition "$\mu^2 > |\sigma^2 - \tau^2|/n$" that characterises good performance for basic SVM and DWD methods, must be strengthened to "$\mu^2 > |\sigma^2 - \tau^2|$" for NND. In these and other ways, the second-order, Cesàro-stationarity model gives theoretical insight into numerical results about the performances of different classifiers.

The model can be altered, and in particular generalised, in a variety of ways, to gain still further information. For example, the way in which the component-wise means and variances change with component index can be adjusted, so that σ^2, τ^2 and μ^2 are all zero, or where for other reasons the marginal cases (e.g., in the example in the previous paragraph, the case $\mu^2 = |\sigma^2 - \tau^2|$) obtain. Furthermore, the distributions of components can be given a degree of heavy-tailed behaviour, or be given stronger dependence, than has been considered in this paper. In these ways, and in others, the simple model suggested here can be used as the basis for a wider range of explorations of the manner in which classifiers compare.

References

[1] ALTER, O., BROWN, P.O. and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy for the Sciences U.S.A.*, 97, 10101-10106.

[2] BAI, Z. and SARANDASA, H. (1996) Effect of high dimension: by an example of a two sample problem, *Statistica Sinica*, 6, 311-329.

[3] BURGES, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 955–974.

[4] COOTES, T.F., HILL, A., TAYLOR, C.J. and HASLAM, J. (1993). The use of active shape models for locating structures in medical images, *Information Processing in Medical Imaging*, H. H. Barret and A. F. Gmitro, eds., Lecture Notes in Computer Science 687, 33-47, Springer Verlag, Berlin.

[5] CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000) *An introduction to Support Vector Machines*, Cambridge University Press.

[6] EISEN, M.B. and BROWN, P.O. (1999) DNA arrays for analysis of gene expression. *Methods for Enzymology*, 303, 179-205.

[7] HUBER, P. J. (1973) Robust regression: asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, 1, 799-821.

[8] JOHNSTONE, I. M. (2001) On the distribution of the largest principal component, *Annals of Statistics*, 29, 295-327.

[9] KOLMOGOROV, A.N. AND ROZANOV, Y.A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory of Probability and its Applications*, 5, 204–208.

[10] MAHALANOBIS, P. C. (1936) On the generalized distance in statistics, *Proceedings of the National Institute of Science, India*, 2, 49-55.

[11] MARRON, J.S. AND TODD, M. (2002) Distance weighted discrimination. Manuscript.

[12] MARRON, J.S., WENDELBERGER, J.R. AND KOBER, E.M. (2004) Time Series Functional Data Analysis, Manuscript.

[13] 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.

[14] PEROU, C.M., SØRLIE, T., EISEN, M.B., VAN DE RIJN, M., JEFFREY, S.S., REES, C. A., POLLACK, J.R., ROSS, D.T., JOHNSEN, H., AKSLEN, L.A., FLUGE, Ø., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S.X., LØNNING, P.E., BØRRESEN-DALE, A.-L., BROWN, P.O. AND BOTSTEIN, D. (2000). Molecular portraits of human breast tumours. *Nature,* 406, 747-52.

[15] PORTNOY, S. (1984) Asymptotic behavior of M–estimators of p regression parameters when p^2/n is large; I. Consistency, *Annals of Statististics*, 12, 1298–1309.

[16] PORTNOY, S. (1988) Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *Annals of Statististics*, 16, 356–366, 1988.

[17] RAO, C. R. and VARADARAJAN, V. S. (1963) Discrimination of Gaussian processes, *Sankhya, Series A*, 25, 303-330.

[18] RAO. C. R. (1973) Mahalanobis era in statistics, Sankhya Ser. B. 35, Part 4, Suppl. 12-36.

[19] SARANDASA, H. and ALTAN, S. (1998) The analysis of small-sample multivariate data, *Journal of Biopharmaceutical Statistics*, 8, 163-186.

[20] SCHÖLKOPF, B. and SMOLA, A. (2001) *Learning with Kernels*, MIT Press, Cambridge, MA.

[21] SCHOONOVER, J. R., MARX, R. and ZHANG, S. L. (2003) Multivariate Curve Resolution in the Analysis of Vibrational Spectroscopy Data Files, *Applied Spectroscopy*, 57, 483-490.

[22] SØRLIE, T., PEROU, C.M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M.B., VAN DE RIJIN, M., JEFFREY, S.S., THORSEN, T., QUIST, H., MATESE, J.C., BROWN, P.H., BOTSTEIN, D., LONNING, P.E., BORRESEN-DALE, A.L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy for the Sciences U.S.A.*, 98, 10869-74.

[23] TSYBAKOV, A. B. (2003) Optimal rates of aggregation, *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines, Lecture Notes in Artificial Intelligence, vol.2777*, Eds. Schölkopf, B. and Warmuth, M., Springer, Heidelberg (2003).

[24] VAPNIK, V.N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer, Berlin.

[25] VAPNIK, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

[26] YUSHKEVICH, P., PIZER, S.M., JOSHI, S. AND MARRON, J.S. (2001) Intuitive localized analysis of shape variability, *Information Processing in Medical Imaging (IPMI)*, eds. Insana, M.F. and Leahy. R.M., 402-408.