J. R. Statist. Soc. B (2005)
67, *Part* 3, *pp.* 427–444

Geometric representation of high dimension, low sample size data

Peter Hall,

Australian National University, Canberra, Australia

J. S. Marron

University of North Carolina, Chapel Hill, USA

and Amnon Neeman

Australian National University, Canberra, Australia

[Received April 2004. Revised February 2005]

Summary. High dimension, low sample size data are emerging in various areas of science. We find a common structure underlying many such data sets by using a non-standard type of asymptotics: the dimension tends to ∞ 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 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.

Keywords: Chemometrics; Large dimensional data; Medical images; Microarrays; Multivariate analysis; Non-standard asymptotics

1. Introduction

High dimension, low sample size (HDLSS) data are becoming increasingly common in various fields. These include genetic microarrays, medical imaging and chemometrics, which we treat briefly in the next three paragraphs.

A currently very active area of data analysis is microarrays 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 analysing populations of three-dimensional images. A useful approach is first to represent numerically shapes of organs of interest. This is done in a wide variety of ways, including the boundary representations that were developed by Cootes *et al.* (1993), and the completely different medial representations, which were 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

Address for correspondence: J. S. Marron, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA.
E-mail: marron@email.unc.edu

 1369–7412/05/67427

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 *et al.* (2003) and Marron *et al.* (2004). As with the above fields, practical considerations limit the number of samples to far fewer than the number of channels, again resulting in $n \ll d$.

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 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 that is 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 this elementary geometric representation.

The asymptotics in this paper treat the HDLSS case of $d \to \infty$, with n fixed, although the case where d and n diverge together, with $d/n^2 \to \infty$, may be addressed similarly. 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 1) 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) discussed 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 analyse the HDLSS performance of some discrimination rules, including the support vector machine (SVM), 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 that are 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 χ^2-distribution with d degrees of freedom, which tends towards the Gaussian distribution as $d \to \infty$, a simple delta method calculation shows that the

Euclidean distance has the property

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

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:

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

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

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

where $\text{ang}(Z_1, Z_2)$ denotes the angle, in measured radians at the origin, between vectors Z_1 and Z_2. Of course, both equations (1) and (2) hold for a random sample $Z_1, \ldots, Z_n$, 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 $n \geqslant 4$.

These properties are illustrated in Fig. 1, where the case $d = 3$ and $n = 3$ is considered. All the rays from the origin to the respective data points are of approximately equal length, and the distances between data points 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 that is generated by the data (of dimension $n = 3$ in this case) or on the hyperplane that is 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 by using the number of dimensions in which it lives, rather than its number of vertices. However, the 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.)

Another example elucidating these ideas is shown in Fig. 2. Each panel shows overlaid scatterplots of 10 samples (shown as different geometrical shapes) of standard normal random vectors

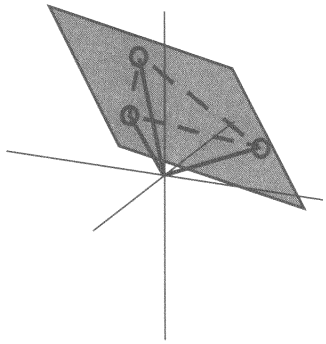

Fig. 1. Three-point toy example, showing the geometric representation, by rotation of the two-dimensional hyperplane containing the data, to give a regular n-hedron

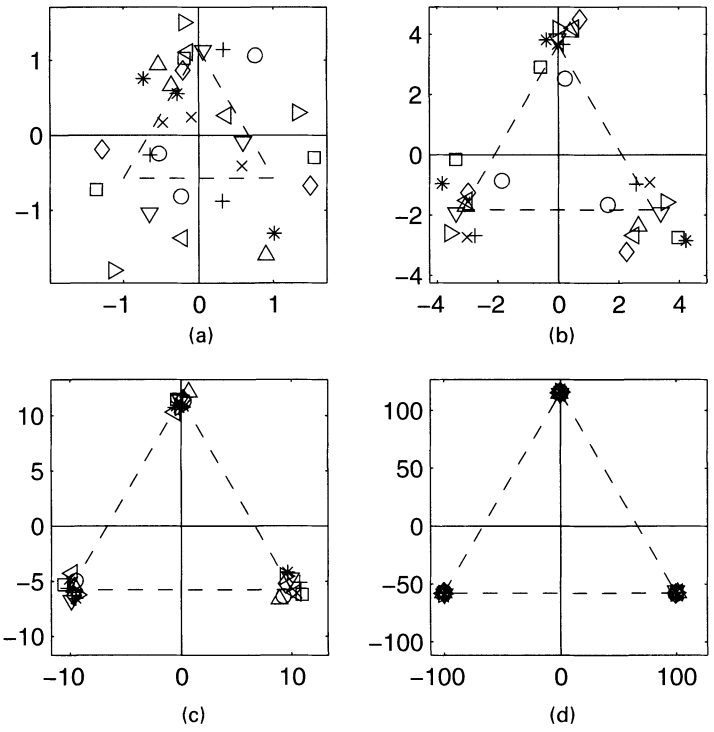

Fig. 2. Gaussian toy example, illustrating the geometric representation, for $n = 3$, and convergence to a 3-simplex with increasing dimension: (a) $d = 2$; (b) $d = 20$; (c) $d = 200$; (d) $d = 20000$

of size $n = 3$, and in $d = 2, 20, 200, 20000$ dimensions in Figs 2(a), 2(b), 2(c) and 2(d) respectively. 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 that is generated by the data (i.e. the plane that is shown in Fig. 1) is found, and the data are projected onto that. Within that hyperplane the data are rotated so that the horizontal co-ordinates of the bottom two points are centred on 0, to give the scatterplots that are shown in Fig. 2. In view of equation (1) it is expected that these points will lie close to the vertices of the equilateral triangle, with side length $(2d)^{1/2}$ shown with the broken lines (the regular 3-simplex}, and that this approximation will be better for higher dimensions.

Fig. 2 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 that were 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 underlie 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 $X(d) = (X^{(1)}, \ldots, X^{(d)})^{\mathrm{T}}$, which is 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

$$d^{-1} \sum_k X^{(k)2} \to a$$

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, which was observed in Section 2, will follow from the limiting behaviour 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 the following.

(a) The fourth moments of the entries of the data vectors are uniformly bounded.
(b) For a constant σ^2,

$$\frac{1}{d} \sum_{k=1}^{d} \mathrm{var}(X^{(k)}) \to \sigma^2. \tag{3}$$

(c) 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

$$\frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} (X_i^{(k)} - X_j^{(k)})^2 \right\}^{1/2} \to (2\sigma^2)^{1/2}, \tag{4}$$

where the convergence is in probability. See Section 5.1 for details on result (4).

Stationarity of the time series X is not required. Instead we need only boundedness of moments, weak independence and condition (3), which entails stationarity only in a very weak, Cesàro-averaged, first-order form. In this sense we are working with a rich class of models for high dimensional data. Assumption (c) is a simple way of permitting the amount of information that is available for discrimination to diverge to ∞ 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 (c) fail; see Section 6.

Application of result (4) to each pair (i, j) with $1 \leqslant i < j \leqslant m$, 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

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

Of course, the theory that is described in conclusion (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 broken edges that is shown in Fig. 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 assumptions (a)–(c) in Section 3.1 for the time series Y, together with a new assumption about separation of population means, gives the new conditions

$$\begin{aligned} &\frac{1}{d}\sum_{k=1}^{d} \mathrm{var}(Y^{(k)}) \to \tau^2, \\ &\frac{1}{d}\sum_{k=1}^{d} \{E(X^{(k)}) - E(Y^{(k)})\}^2 \to \mu^2, \end{aligned} \tag{6}$$

where τ and μ denote finite positive constants. It follows that the analogue of result (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 conditions (6) is especially relevant to accurate classification. If μ in expression (6) is too small, and in particular if it equals 0, then a classifier of any conventional type (SVM, distance-weighted discrimination (DWD), 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, i.e. 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 $d \to \infty$:

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

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 N-polyhedron that is discussed at result (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.

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

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

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

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 disappear 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 0, 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 behaviour of discrimination methods. In particular, in the simulation study of Marron and Todd (2005), it was observed that, at very high dimensions, the techniques considered all had similar error rates, across a wide array of simulation settings. A basic version of the popular SVM and the more recently developed DWD method 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 to whether 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 $d \geqslant N$, and no k data points lie in a $(k-2)$-dimensional hyperplane (which happens with probability 1 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 shall study how the different classification methods vary in terms of the hyperplane that they select.

The SVM method (see for example Vapnik (1982, 1995), Burges (1998), Christianini and Shawe-Taylor (2000) and Schölkopf and Smola (2001)) has been implemented and studied in a wide variety of 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 data sets. These points do not have to be data values. In the asymptotic geometric representation that is described at result (8), these convex hulls are precisely the m- and n-simplices, the vertices of which represent the limits, as $d \to \infty$, of the data sets $\mathcal{X}(d)$ 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 that is 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 make the following claim.

Theorem 1. Assume that $\sigma^2/m \geqslant \tau^2/n$; 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 that was 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 an asymptotically correct classification of new X-values whenever m is sufficiently large, for any given value of n, and an 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 $\mu^2/\sigma^2 < |m^{-1} - n^{-1}|$, 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 (2005) 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

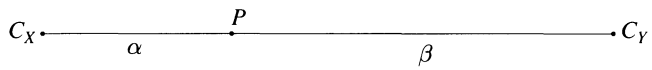

Fig. 3. Relative relationships of simplex centroids C_X, C_Y and the candidate DWD cut-off point P

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 $\mathcal{X}(d)$ and 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 that are 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 interval $C_X C_Y$. We want to see when it lies on the DWD hyperplane. Relative relationships are diagrammed in Fig. 3.

Because the simplex $\mathcal{X}(d)$ is orthogonal to $C_X C_Y$, all the vertices in the simplex are distance α from the hyperplane passing through P, orthogonal to $C_X C_Y$. 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

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

This tells us the location of the DWD hyperplane. It is the hyperplane that is orthogonal to the line $C_X C_Y$, passing through the point P which satisfies condition (9). In Section 5.3 we shall 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 that 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 we have the following.

Theorem 2. Assume that $\sigma^2/m^{(p+2)/(p+1)} \geqslant \tau^2/n^{(p+2)/(p+1)}$; 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.

As $p \to \infty$, theorems 1 and 2 become identical. More generally, the rules which determine success or failure of classification, using the basic SVM or DWD, are similar when p is large. In this sense, the basic SVM can be viewed as a limiting case of DWD; the 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. In contrast, 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 that was observed in Fig. 5 of Marron and Todd (2005), that for spherical Gaussian data DWD gave a somewhat better classification performance than the basic SVM. However, in other cases, in particular where the conditions of the theorems are not well preserved, the SVM can outperform DWD. See the end of Section 4.3 for an example.

4.2. Other discrimination rules

Let $C_X(d)$ and $C_Y(d)$ denote the centroids of the data sets $\mathcal{X}(d)$ 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 to whether Z is closer to $C_X(d)$ or to $C_Y(d)$ 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, i.e. 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 the CRD is driven by averaging the stochastic errors, not by the extrema. This is a new way of understanding the superior performance of the CRD over the basic SVM in the example that was considered in Fig. 5 of Marron and Todd (2005). DWD gave essentially the same performance in that case because the sample sizes were equal.

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

Theorem 3. Assume that $\sigma^2 \geqslant \tau^2$; 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, the basic SVM misclassifies data from at least one of the populations only when $\mu^2 < |\sigma^2 - \tau^2|/m$, whereas theorem 3 asserts that the NND leads to misclassification, for data from at least one of the populations, both in this range and when $|\sigma^2 - \tau^2|/m \leqslant \mu^2 < |\sigma^2 - \tau^2|$. This quantifies the inefficiency that might be expected from basing inference on only a single nearest neighbour. Furthermore, without the condition

$m = n$, the basic SVM has an asymptotic advantage over the NND, in the sense of leading to the 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, if the CRD and DWD (for $m = n$, or for large p) classifiers are equivalent to the basic SVM, then the remarks in the previous paragraph remain true if we replace the basic SVM by either DWD or CRD. This explains the observation of Marron and Todd (2005) that these methods all gave similar simulation results for very large dimension d (in the case of $m = n$). Furthermore, the four classifiers that are considered here divide naturally into two groups. The first group contains the 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 the 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 by 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 that are 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 (2005) was that in a variety of simulation settings considered there, for all of the basic SVM, DWD and CRD, the classification error rates tended to come together for large

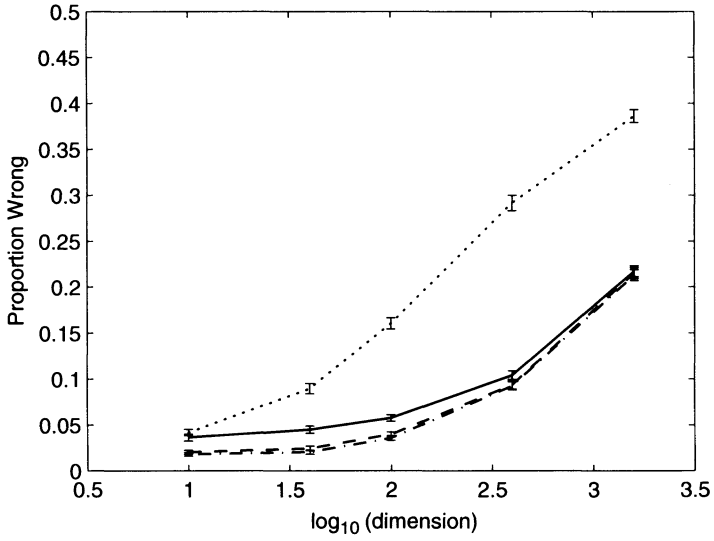

Fig. 4. Summary of the simulation results, for Gaussian data, showing convergence of most of the methods for large dimension: ———, SVM; – – –, DWD; · – · – ·, CRD; · · · · · · ·, NND

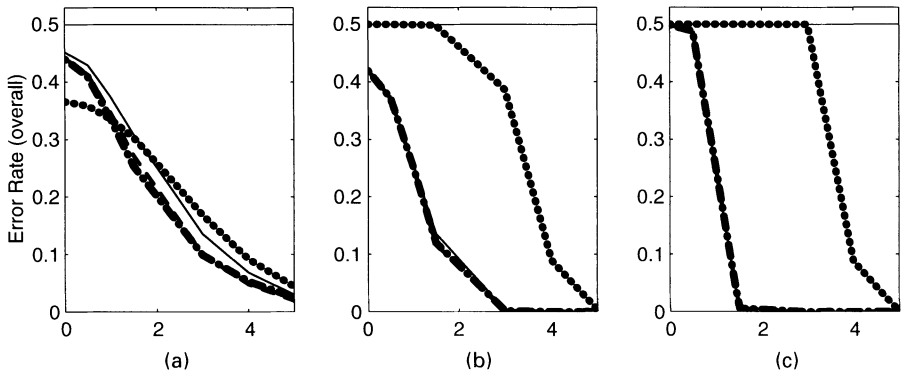

Fig. 5. Summary of simulations exploring asymptotic lessons, showing 'changepoints' at predicted values of μ (———, SVM; – – –, DWD; · – · – ·, CRD; · · · · · ·, NND): (a) $d = 10$; (b) $d = 100$; (c) $d = 1000$

d. Fig. 4 is similar to Fig. 5 of Marron and Todd, except that the NND has now been added. This shows overall error rates, for the four classification methods that are 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 $X_i^{(1)}$, $i = 1, \ldots, m$ (and of $Y_j^{(1)}$ $j = 1, \ldots, n$), has been shifted to 2.2 (and -2.2 respectively). Classification error rates were computed on the basis of 100 new data points from each of the two classes, and the means are summarized by the curves in Fig. 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 entries 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 the 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 the basic SVM, DWD and CRD all converge to 1 as d increases.) Finally, again as predicted, the basic SVM lags somewhat behind DWD and the CRD (which are not significantly different).

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

The ideas of theorems 1–3 are illustrated in a different way in Fig. 5. The simulation setting of Fig. 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, $d = 10, 100, 1000$, are shown in Figs 5(a), 5(b) and 5(c) respectively. The classification methods are distinguished by using different line types. Different line thicknesses are used to decrease overplotting effects; for example, for $d = 1000$, the basic SVM, DWD and CRD results are essentially on top of each other. Again error rates are computed using 100 new test cases for each class, and averaged over 1000 Monte Carlo repetitions.

Fig. 5 allows a convenient study of the classification error rate as a function of μ. Theorem 1 suggests that for 'μ large' perfect discrimination (i.e. error rate 0) is possible for the basic SVM, which is reflected by the full curves coming to 0 on the right-hand 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 changepoint at $\mu = (\sigma^2/m - \tau^2/n)^{1/2} = (20/16 - 4/16)^{1/2} = 1$ expected. To the left of the changepoint, 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 changepoint 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 is seen in Fig. 5 as the broken curve. The performance is virtually identical to that of the basic SVM for $d = 1000$ and, again as predicted at the end of Section 4.1, DWD is marginally better for $d = 10$ and $d = 100$.

Recall from theorem 3 that for NND the changepoint is quite different, now appearing for $\mu = (\sigma^2 - \tau^2)^{1/2} = (20 - 4)^{1/2} = 4$ (further to the right, reflecting the expected inefficiency of one nearest neighbour discrimination). This changepoint is also well reflected in Fig. 5, as the curves comprised of bold dots. 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 Fig. 6. Specific results are shown for the case of $m = 2$ and $n = 5$. Fairly similar results were obtained for other values of m and n.

As in Fig. 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 the line type 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 $E(X^{(k)}) = 0$ and $E(Y^{(k)}) = \mu$, and the variances were taken to be $\sigma^2 = \tau^2 = 1$. Again error rates are displayed as a function of

$$\mu = \left[\frac{1}{d}\sum_{k=1}^{d}\{E(X^{(k)}) - E(Y^{(k)})\}^2\right]^{1/2}.$$

Once again the lessons from asymptotic prediction 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 full curves). The $n = 5$ class $\mathcal{Y}(d)$ error rates (indicated by medium

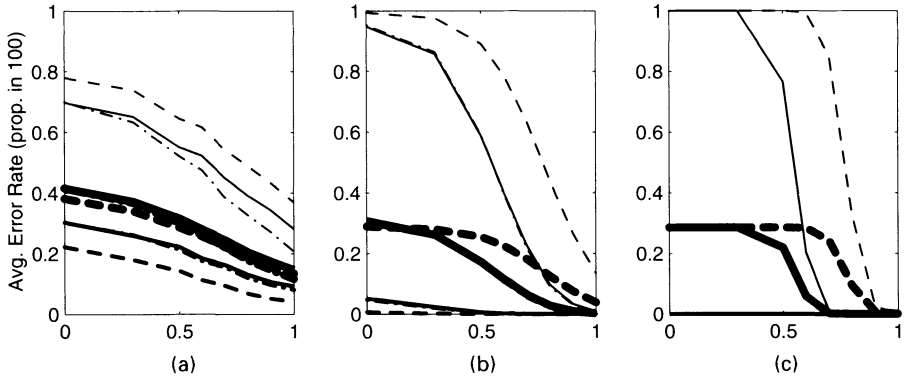

Fig. 6. Summary of simulations exploring asymptotic lessons for unequal sample sizes $m = 2$ and $n = 5$, showing 'changepoints' at predicted values of μ (———, SVM; – – –, DWD; · – · – ·, CRD; ———, class X; ———, class Y; ▬▬, overall): (a) $d = 10$; (b) $d = 100$; (c) $d = 1000$

width lines) are much smaller and decrease to all 0 for $d = 1000$ (the medium full 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 the rates decrease, with the fastest decrease for $d = 1000$ (shown by using full curves).

Also as predicted, for the highest $d = 1000$, the CRD and SVM give essentially the same results (i.e. for each line thickness the chain curve is always almost on top of the corresponding full curve), and DWD is substantially worse in terms of both class $\mathcal{X}(d)$ and overall error rates (the thin and thick broken curves are higher than the corresponding curves of other line types). For lower dimensions, the results are fairly similar, but DWD seems to give better class $\mathcal{Y}(d)$ performance (shown by the dashed medium width lines being below the other two line types), 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 thick broken curve being below the other thick curves).

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

5. Technical details

This section gives the technical details that were 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 $|f(u, v)| + |g(u, v)| \leqslant Cu^2v^2$ for fixed $C > 0$ and all u and v, we have

$$\sup_{1 \leqslant k,l < \infty, |k-l| \geqslant r} |\mathrm{corr}\{f(U^{(k)}, V^{(k)}), g(U^{(k)}, V^{(k)})\}| \leqslant \rho(r),$$

for $(U, V) = (X, X), (Y, Y), (X, Y)$, 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,

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

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

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

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

5.2. Derivation for basic support vector machine

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 that is 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 co-ordinates $(1,0,\ldots,0),\ldots,(0,\ldots,0,1)$. (This m-variate representation is simpler than an $(m-1)$-variate representation.) Then the centroid of the simplex has co-ordinates $(m^{-1},\ldots,m^{-1})$, 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 that was 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

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

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 that $\sigma^2/m > \tau^2/n$. 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 the information about data points of type X and Y.

5.3. *Derivation for distance-weighted discrimination*

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

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

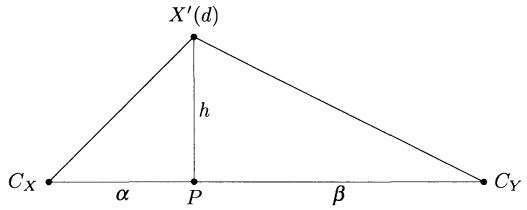

Fig. 7. Relative relationships between the new datum point $X'(d)$ and the simplex centroids C_X and C_Y

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 that is shown in Fig. 7, we know the distances $C_X C_Y$, $X'(d)C_X$ and $X'(d)C_Y$.

In Fig. 7, P is the projection of $X'(d)$ to the line $C_X C_Y$. The distances that 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 equation (11) from equation (10) we have

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

Adding equations (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, i.e. if

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

It will be classified as belonging to Y if

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

So far our treatment has been general. Now assume that

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

The analysis above tells us when a point $X'(d)$ will be classified correctly. Suppose that we have a point $Y'(d)$. By the inequality above

$$\frac{\tau^2/n}{\sigma^2/m} \leqslant \left(\frac{n}{m}\right)^{1/(p+1)}.$$

But then for any positive μ^2 we have

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

i.e. $Y'(d)$ 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 nearest neighbour discriminator

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 normalization, $X'(d)$ converges to a point whose squared distances 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 to whether $2\sigma^2 < l^2$ or $2\sigma^2 > l^2$ respectively. Since $2\sigma^2 < l^2$ if and only if $\mu^2 > \sigma^2 - \tau^2$, theorem 3 follows.

6. Summarizing remarks 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 expressions (3) and (6)), the performances of different classifiers for very high dimensions can be represented, quite simply, in terms of the relationships between the average co-ordinate variances, and the average squared differences of means. This analysis has revealed a variety of properties of different classifiers. For example, it has been shown that the basic SVM and DWD classifiers perform similarly when the sample sizes are the same, but not necessarily when the sample sizes differ, and that, from some perspectives, the 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 the 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 characterizes good performance for the basic SVM and DWD methods must be strengthened to '$\mu^2 > |\sigma^2 - \tau^2|$' for the 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 generalized, in a variety of ways, to gain still further information. For example, the way in which the componentwise means and variances change with the component index can be adjusted, so that σ^2, τ^2 and μ^2 are all 0, 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 that is suggested here can be used as the basis for a wider range of explorations of the manner in which classifiers compare.

References

Alter, O., Brown, P. O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natn. Acad. Sci. USA*, **97**, 10101–10106.

Bai, Z. and Sarandasa, H. (1996) Effect of high dimension: by an example of a two sample problem. *Statist. Sin.*, **6**, 311–329.

Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Minng Knowl. Disc.*, **2**, 955–974.

Cootes, T. F., Hill, A., Taylor, C. J. and Haslam, J. (1993) The use of active shape models for locating structures in medical images. *Lect. Notes Comput. Sci.*, **687**, 33–47.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.
Eisen, M. B. and Brown, P. O. (1999) DNA arrays for analysis of gene expression. *Meth. Enzym.*, **303**, 179–205.
Huber, P. J. (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.
Johnstone, I. M. (2001) On the distribution of the largest principal component. *Ann. Statist.*, **29**, 295–327.
Kolmogorov, A. N. and Rozanov, Y. A. (1960) On strong mixing conditions for stationary Gaussian processes. *Theory Probab. Appl.*, **5**, 204–208.
Mahalanobis, P. C. (1936) On the generalized distance in statistics. *Proc. Natn. Inst. Sci. Ind.*, **2**, 49–55.
Marron, J. S. and Todd, M. (2005) Distance weighted discrimination. *J. Am. Statist. Ass.*, to be published.
Marron, J. S., Wendelberger, J. R. and Kober, E. M. (2004) Time series functional data analysis. *Report LA-UR-04-3911*. Los Alamos National Laboratory, Los Alamos.
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. *Proc. Natn. Acad. Sci. USA*, **96**, 9212–9217.
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**, 747–752.
Portnoy, S. (1984) Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; I, consistency. *Ann. Statist.*, **12**, 1298–1309.
Portnoy, S. (1988) Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, **16**, 356–366.
Rao, C. R. (1973) Mahalanobis era in statistics. *Sankhya* B, **35**, suppl., 12–36.
Rao, C. R. and Varadarajan, V. S. (1963) Discrimination of Gaussian processes. *Sankhya* A, **25**, 303–330.
Sarandasa, H. and Altan, S. (1998) The analysis of small-sample multivariate data. *J. Biopharm. Statist.*, **8**, 163–186.
Schölkopf, B. and Smola, A. (2001) *Learning with Kernels*. Cambridge: MIT Press.
Schoonover, J. R., Marx, R. and Zhang, S. L. (2003) Multivariate curve resolution in the analysis of vibrational spectroscopy data files. *Appl. Spectrosc.*, **57**, 483–490.
Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. H., Botstein, D., Lonning, P. E. and Borresen-Dale, A. L. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natn. Acad. Sci. USA*, **98**, 10869–10874.
Tsybakov, A. B. (2003) Optimal rates of aggregation. *Lect. Notes Artif. Intell.*, **2777**.
Vapnik, V. N. (1982) *Estimation of Dependences based on Empirical Data*. Berlin: Springer.
Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. New York: Springer.
Yushkevich, P., Pizer, S. M., Joshi, S. and Marron, J. S. (2001) Intuitive localized analysis of shape variability. In *Information Processing in Medical Imaging* (eds M. F. Insana and R. M. Leahy), pp. 402–408.