J. R. Statist. Soc. B (2008)
70, *Part* 1, *pp.* 159–173

Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes

Peter Hall,

Australian National University, Canberra, and University of Melbourne, Australia

Yvonne Pittelkow

Australian National University, Canberra, Australia

and Malay Ghosh

University of Florida, Gainesville, USA

[Received September 2006. Revised August 2007]

Summary. We suggest a technique, related to the concept of 'detection boundary' that was developed by Ingster and by Donoho and Jin, for comparing the theoretical performance of classifiers constructed from small training samples of very large vectors. The resulting 'classification boundaries' are obtained for a variety of distance-based methods, including the support vector machine, distance-weighted discrimination and *k*th-nearest-neighbour classifiers, for thresholded forms of those methods, and for techniques based on Donoho and Jin's higher criticism approach to signal detection. Assessed in these terms, standard distance-based methods are shown to be capable only of detecting differences between populations when those differences can be estimated consistently. However, the thresholded forms of distance-based classifiers can do better, and in particular can correctly classify data even when differences between distributions are only detectable, not estimable. Other methods, including higher criticism classifiers, can on occasion perform better still, but they tend to be more limited in scope, requiring substantially more information about the marginal distributions. Moreover, as tail weight becomes heavier the classification boundaries of methods designed for particular distribution types can converge to, and achieve, the boundary for thresholded nearest neighbour approaches. For example, although higher criticism has a lower classification boundary, and in this sense performs better, in the case of normal data, the boundaries are identical for exponentially distributed data when both sample sizes equal 1.

Keywords: Classification boundary; Detection; Distance-based classification; Distance-weighted discrimination; Higher criticism; Nearest neighbour method; Sparsity; Support vector machine; Thresholding; Truncation

1. Introduction

A variety of classification methods have been developed in response to discrimination problems that are posed by small sample sizes, sparsity and high dimensional data. They include a range of distance-based classifiers, such as the support vector machine, distance-weighted discrimination and *k*th-nearest-neighbour techniques. In this paper we suggest an approach to assessing

Address for correspondence: Peter Hall, Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010 Australia.
E-mail: P.Hall@ms.unimelb.edu.au

 1369–7412/08/70159

the theoretical performance of such methods, and for comparing them with their counterparts based on thresholding, and with alternative classifiers that are inspired by higher criticism ideas.

We shall show that each member of a large class of distance-based classifiers has the same classification boundary. Likewise, thresholded versions of these methods share the same boundary, although now the envelope lies strictly below its counterpart for non-thresholded classifiers, i.e. the thresholding methods perform better than their non-thresholded counterparts, provided that the threshold is chosen appropriately. The non-thresholded classifiers can use effectively only those distribution differences that can be estimated, but, through gains in reduction of noise, their thresholded forms can exploit differences that are too small for estimation.

A different class of classifiers, including higher criticism approaches, can be developed from sparse signal detection methods that were discussed by Ingster (1999, 2001, 2002) and Donoho and Jin (2004). In some instances the classification boundary for higher criticism is lower still than that for thresholded, distance-based classifiers, although the higher criticism method is restricted to problems where relatively good information about marginal distributions, and near independence among marginals, is available. Moreover, the classification boundary for methods that are based on higher criticism converges to that for thresholded nearest neighbour methods as the tail weight of marginal distributions increases.

One implication of our results is that classification boundaries for distance-based classifiers, and for their thresholded forms, do not depend on the fixed training sample sizes. This conclusion is a consequence of the fact that, in the fixed sample size case and for distance-based classifiers, the probability of correct classification converges to 1 if and only if the (squared) differences between distances among data values have a certain extremal property; and that this property holds for one difference if and only if it holds for all of them. See equation (3.11) for an example of the sort of difference to which we are referring. Therefore it does not matter how many distances, or, equivalently, how many training data, there are. (Here we are referring to the case of fixed sample size.) However, for other types of classifier, which depend on information about the marginal distributions, the sizes of the training samples can affect the formula for the classification boundary. In such instances, the case where both training sample sizes equal 1 is a convenient benchmark.

Sparsity, which makes classification of very high dimensional data particularly challenging, results in the information that can assist classification being available only at a very small proportion of components, scattered through a particularly long data vector. The classifier must implicitly identify the unknown locations of this information, and use the information there effectively. The first of these two challenges typically does not arise in conventional classification problems, where the number of components is much less than the sample size. Examples of real situations where sparsity arises include those which were addressed by Meinshausen and Rice (2004), where, out of 10^{11} vector components in the study of the Kuiper Belt in astronomy, only a few hundred components potentially contain the signal, and by Efron (2004), where fewer than 20 out of several thousand genes are differentially expressed.

Related research on higher criticism and related techniques for sparse signal detection includes contributions by Jin (2002, 2005, 2006) and Cayon *et al.* (2005, 2006). Support vector machine methods enjoy a wide variety of generalizations and modifications; see, for example, Vapnik (1982, 1995), Burges (1998), Brown *et al.* (2000), Cristianini and Shawe-Taylor (2000) and Schölokopf and Smola (2001). However, in cases where the sample sizes are much less than the dimension, there are relatively few opportunities for altering support vector machine approaches. The definition that we shall consider is the classical one.

Distance-weighted discrimination has been proposed and explored by Marron *et al.* (2005), and nearest neighbour methods have been described by, for example, Murtagh (1985) and Choi

and Rockett (2002). Asymptotic statistical properties, in problems where sample sizes and data vector length both increase, have been discussed by Bai and Sarandasa (1996), Sarandasa and Altan (1998), Johnstone (2001), Baik *et al.* (2005) and El Karoui (2005), among others. Hall *et al.* (2005) have treated problems where sample size is kept fixed and vector length is permitted to increase, although the properties that were treated there are quite different from those discussed here.

2. Methods for classification

2.1. Distance-based classifiers

Given data, in the form of samples of p-vectors $\mathcal{X} = \{X_1, \ldots, X_m\}$ from population Π_X and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ from population Π_Y, we wish to allocate a new observation, Z say, to either Π_X or Π_Y. We define distance among p-vectors in the usual Euclidean way.

Let $\mathcal{C} = \mathcal{C}(\cdot|\mathcal{X}, \mathcal{Y})$ denote a classifier which assigns Z to either Π_X or Π_Y, meaning that, with probability 1, either $\mathcal{C}(Z) = \Pi_X$ or $\mathcal{C}(Z) = \Pi_Y$. We argue that any plausible, distance-based classifier $\mathcal{C}$ should enjoy both of the following properties:

(a) $\mathcal{C}$ classifies Z as coming from Π_X if it is closer to each of the X_is than it is to any of the Y_is;
(b) if $\mathcal{C}$ classifies Z as coming from Π_X then at least one of the X_is is closer to Z than Z is to the most distant Y_i.

Of course, property (b) merely states that the analogue of property (a) applies when classifying data as coming from Π_Y. Standard classification rules which have these properties include the support vector machine, distance-weighted discrimination and kth-nearest-neighbour classifiers.

Together, properties (a) and (b) imply that

$$\pi_{W1} \leqslant P_W\{\mathcal{C}(Z) = \Pi_X\} \leqslant \pi_{W2} \qquad \text{for } W = X, Y, \tag{2.1}$$

where P_W denotes probability measure under the assumption that Z is from population Π_W, and π_{W1} and π_{W2} are defined by

$$\begin{aligned} \pi_{W1} &= P_W(\max_{1 \leqslant i \leqslant m} \|X_i - Z\| \leqslant \min_{1 \leqslant i \leqslant n} \|Y_i - Z\|), \\ \pi_{W2} &= P_W(\min_{1 \leqslant i \leqslant m} \|X_i - Z\| \leqslant \max_{1 \leqslant i \leqslant n} \|Y_i - Z\|). \end{aligned} \tag{2.2}$$

All the distance-based classifiers that we shall consider will be assumed to satisfy condition (2.1).

In the theory that is developed in Section 3 we shall permit dimension p to increase, while keeping training sample sizes m and n fixed. This reflects a range of commonly occurring problems, where m and n range between 1 and 10 whereas p is many hundreds, or many thousands, in size. In such cases, information for classification accumulates through a large number of vector components rather than a large number of data values.

2.2. Thresholding methods for increasing classifier sensitivity

To improve the sensitivity of classifiers it is common to threshold below at relatively high positive values, or high absolute values, to emphasize vector components that are believed to have high leverage for classification.

In particular, with X_{ij}, Y_{ij} and Z_j denoting the jth components of X_i, Y_i and Z respectively, for $1 \leqslant j \leqslant p$, and with $t > 0$ representing a threshold which generally depends on p, we threshold X_{ij}, Y_{ij} and Z_j, or their absolute values, at t. We treat in detail only the setting where the components themselves are thresholded; the case of absolute values is similar. Therefore, we replace X_{ij}, Y_{ij} and Z_{ij} by $X_{ij}^{\rm tr} = X_{ij}\, I(X_{ij} > t)$, $Y_{ij}^{\rm tr} = Y_{ij}\, I(Y_{ij} > t)$ and $Z_j^{\rm tr} = Z_j\, I(Z_j > t)$ respectively.

The corresponding thresholded vectors are $X_i^{\rm tr} = (X_{ij}^{\rm tr})$, $Y_i^{\rm tr} = (Y_{ij}^{\rm tr})$ and $Z^{\rm tr} = (Z_j^{\rm tr})$. Let $\mathcal{C}^{\rm tr}$ denote the version of the classifier $\mathcal{C}$ that arises when the latter is applied to the thresholded data sets $\mathcal{X}^{\rm tr} = \{X_1^{\rm tr}, \ldots, X_m^{\rm tr}\}$ and $\mathcal{Y}^{\rm tr} = \{Y_1^{\rm tr}, \ldots, Y_n^{\rm tr}\}$, instead of to the original data in $\mathcal{X}$ and $\mathcal{Y}$.

This approach to thresholding replaces a data value that is less than t by 0. That is sometimes done in practice, but more commonly it is removed altogether. However, in terms of the theoretical results that we shall derive, removal is equivalent to replacement by 0. Empirical choice of t will be discussed in Section 4.3.

2.3. Classifiers based on higher criticism

Although, as we shall show in Section 3, $\mathcal{C}^{\rm tr}$ has the potential to outperform $\mathcal{C}$ in terms of sensitivity, it can still perform less well than classifiers that exploit knowledge of tail behaviour. For sparse, normally distributed data, classifiers can be developed from suggestions that were made by Ingster (1999) and Donoho and Jin (2004) for detecting small levels of contamination. One approach can be founded on Tukey's notion of 'higher criticism', by conducting, for each j, a test of significance using the jth component of Z and the vectors in $\mathcal{X} \cup \mathcal{Y}$. The resulting classifier is based on the statistical significance of the number of statistically significant results among these tests.

Since higher criticism requires significant information about component distributions then, for definiteness, we shall assume that the data X_{ij} and Y_{ij} are independent $\mathrm{GN}_\gamma(\mu_{Xj}, \sigma_X^2)$ and $\mathrm{GN}_\gamma(\mu_{Yj}, \sigma_Y^2)$ respectively, where $\mathrm{GN}_\gamma(\mu, \sigma^2)$ denotes the Subbotin, or generalized normal, distribution with probability density

$$f(x|\gamma, \mu, \sigma) = C_\gamma \sigma^{-1} \exp\left(-\frac{|x-\mu|^\gamma}{\gamma\sigma^\gamma}\right),$$

with $\gamma, \sigma > 0$, $-\infty < \mu < \infty$ and $C_\gamma^{-1} = 2\,\Gamma(1/\gamma)\gamma^{1/\gamma - 1}$. Here, μ and σ^2 denote the mean and variance respectively of the distribution $\mathrm{GN}_\gamma(\mu, \sigma^2)$. Donoho and Jin (2004) recounted the interest in, and potential applications of, Subbotin distributions. Of course, the standard normal distribution is just the standard Subbotin distribution $\mathrm{GN}_\gamma(0, 1)$ with $\gamma = 2$. The quantile of the $\mathrm{GN}_\gamma(\mu, \sigma^2)$ distribution corresponding to an upper tail probability of $p^{-\beta}$ equals $\{1 + o(1)\}\{\gamma\beta\, \log(p)\}^{1/\gamma}\sigma$; in Section 4.1 we shall take $\frac{1}{2} < \beta < 1$.

Supposing for the present that σ_X^2 and σ_Y^2 are known, write Φ_γ and $\Psi_\gamma = 2\Phi_\gamma - 1$ for respectively the distribution function of a standard Subbotin variable and of the absolute value of that variable. Let W denote either X or Y, and put $n_X = m$ and $n_Y = n$. Let

$$\rho_{Wj} = \Psi_\gamma\left[\frac{|Z_j - \bar{W}_{.j}|}{\{(1 + n_W^{-1})\sigma_W^2\}^{1/2}}\right],$$

where

$$\bar{X}_{.j} = m^{-1}\sum_i^m X_{ij}$$

and $\bar{Y}_{\cdot j}$ is defined analogously. Order the values ρ_{Wj} as $\rho_{W,(1)} \leqslant \ldots \leqslant \rho_{W,(p)}$, and put

$$\mathrm{hc}_W = p^{1/2} \min_{1 \leqslant j \leqslant p} \left[\frac{jp^{-1} - \rho_{W,(j)}}{\{\rho_{W,(j)}(1 - \rho_{W,(j)})\}^{1/2}} \right]. \tag{2.3}$$

The higher criticism classifier $\mathcal{C}^{\mathrm{hc}}$ allocates Z to whichever population Π_W, for $W = X$ or $W = Y$, has the larger value of hc_W.

The assumption that the variances σ_X^2 and σ_Y^2 are known can be removed by using empirical approximations to those quantities; see the last paragraph of this section. When, in the assumption of $\mathrm{GN}_\gamma(\mu, \sigma^2)$ distributions for X_{ij} and Y_{ij}, we take $\gamma = 2$ (i.e. when we assume normality), the constraint can be relaxed by using moderate deviation arguments to address more general cases. For example, in practice it is not uncommon for the variables X_{ij} and Y_{ij} to represent the values of t-statistics computed from samples, in which case the central limit theorem may ensure approximate normality.

More difficult to mitigate, in the context of higher criticism, is the assumption that the distribution of X_{ij}, for example, does not depend on j, modulo a change of location. Although this condition will be imposed in the mathematical model in Section 3.1, it is not crucial to the performance of distance-based classifiers since those methods are implicitly founded on averages over the index j. Likewise, in the context of distance-based methods, central limit theory for mixtures, with a sufficiently fast but polynomial mixing rate (see for example Ibragimov (1962) and Politis *et al.* (1997)), is readily used to weaken the independence assumption. In the case of higher criticism methods, however, it seems necessary to assume mixing at an exponential rate. All these issues reflect the practical advantages that the classifier $\mathcal{C}^{\mathrm{tr}}$ has over $\mathcal{C}^{\mathrm{hc}}$, although the latter is of substantial interest since, when $m = n = 1$, it delineates performance in the benchmark case of approximately independent and identically distributed components. Indeed, when $m = n = 1$ the arguments of Donoho and Jin (2004) can be used to show that $\mathcal{C}^{\mathrm{hc}}$ has optimal performance when the noise distribution is known to be Gaussian.

Classification techniques that are based on thresholding of distance-based methods, and higher criticism classification, are similar in that both are founded on a form of truncation. This might not be immediately apparent from the definition of hc_W at equation (2.3). However, the right-hand side of that formula can be written almost equivalently as

$$\inf_{c_1 \leqslant t \leqslant c_2} \left(\frac{\sum_j I_{Wj}(t) - p\, \Psi_\gamma(t)}{[p\, \Psi_\gamma(t)\{1 - \Psi_\gamma(t)\}]^{1/2}} \right),$$

where $I_{Wj}(t)$ equals 1 if $|Z_j - \bar{W}_{\cdot j}| / \{(1 + n_W^{-1})\sigma_W^2\}^{1/2} \leqslant t$ and 0 otherwise, and $0 < c_1 < c_2 < \infty$ and c_1 and c_2, depending on n, are chosen sufficiently small and sufficiently large respectively. Thus, there is a sense in which hc_W is based on a continuum of truncation operations at thresholds t. Both higher criticism and distance-based classifiers involve accumulating the effects of data truncation, but in the case of higher criticism the truncation operations are many fold, and are used in a more precise way, with less opportunity for influence from noise.

Variance estimation is simplest when both m and n are at least 2. There, estimators can be based on pairwise differences, e.g. $X_{i_1 j} - X_{i_2 j}$, from which any mean effects cancel. The role of one of the data vectors in this difference could be played by Z, although there is clearly potential for bias to be introduced at this point.

3. Properties of classifiers

3.1. Model for data

Given a classifier satisfying condition (2.1), we shall discuss properties of the probability of mistakenly classifying a data vector Z from Π_Y as coming from Π_X. To model the distributions corresponding to Π_X and Π_Y, assume that we may write

$$\left.\begin{aligned} X_{ij} &= \delta_{ij}, \\ Y_{ij} &= \mu_j + \varepsilon_{ij}, \\ Z_j &= \mu_j + \delta_j, \end{aligned}\right\} \tag{3.1}$$

where the errors δ_{ij}, ε_{ij} and δ_j are all independent and identically distributed as the random variable δ, say, with zero mean and finite fourth moment, and the μ_js are deterministic.

We mention again a point that was made earlier: in the case of the distance-based classifiers $\mathcal{C}$ and $\mathcal{C}^{\rm tr}$, the assumption that components are independent and that the distributions of δ_{ij}, ε_{ij} and δ_j do not depend on j, are inessential. In the context of $\mathcal{C}^{\rm hc}$, however, they are relatively difficult to relax. Further discussion is given in Sections 3.2 and 3.3, where we discuss $\mathcal{C}$ and $\mathcal{C}^{\rm tr}$ respectively.

We keep the distribution of δ fixed throughout our analysis, but we allow the μ_js to vary with j and, although not indicated in notation, also to vary with vector length p. This enables us to make the classification problem more difficult as information is accumulated while p increases. Note also that π_{W1} and π_{W2}, which are defined at expression (2.2), are invariant under any change of location for both populations Π_X and Π_Y, and so the assumption that one of the populations, chosen at expression (3.1) to be Π_X, has zero mean is made without loss of generality.

3.2. Properties of standard distance-based classifiers $\mathcal{C}$

We shall show that, under the model that is described in Section 3.1, and assuming the negligibility condition (3.5) below,

> the probability that the classifier $\mathcal{C}$ correctly classifies a new data value from Π_X or Π_Y converges to 1 if and only if $p = o(\|\mu\|^4)$ as $p \to \infty$, (3.2)

where $\mu = (\mu_1, \ldots, \mu_p)$ and $\|\mu\|^2 = \Sigma_j \mu_j^2$. Recall from Section 2.1 that $\mathcal{C}$ can be quite general, e.g. based on the support vector machine, distance-weighted discrimination or kth-nearest-neighbour classifiers. In particular, property (3.2) holds in this general context.

Property (3.2) gives a concise asymptotic description of the performance of the classifier $\mathcal{C}$. In particular, it tells us just how fast the norm of the mean vector μ must grow for it to be possible to distinguish between Π_X and Π_Y. Our theoretical comparison, in Section 4.1, of different classifiers will be based on assessing their respective performance characteristics, expressed similarly to property (3.2).

Property (3.2) remains true without the assumption that the disturbances δ_{ij}, ε_{ij} and δ_j all have the same distribution. In particular, it holds if this condition is replaced by the assumption that the variances of δ_{ij}, ε_{ij} and δ_j are uniformly bounded away from zero, and that their $2c$th moments are uniformly bounded, where c is as in condition (3.5) below.

To derive property (3.2), note that

$$\|X_{i_1} - Z\|^2 - \|Y_{i_2} - Z\|^2 = V_{i_1 i_2} + \|\mu\|^2, \tag{3.3}$$

where

$$V_{i_1 i_2} = \sum_{j=1}^{p} \{\delta_{i_1 j}^2 - \varepsilon_{i_2 j}^2 - 2\delta_{i_1 j}\mu_j - 2(\delta_{i_1 j} - \mu_j - \varepsilon_{i_2 j})\delta_j\}. \tag{3.4}$$

It will be shown in Section 5.1 that, provided that $E|\delta|^{2c} < \infty$ for some $c \geqslant 2$, and

$$\max(p, \|\mu\|^2)^{-c} \sum_{j=1}^{p} |\mu_j|^{2c} \to 0 \tag{3.5}$$

as $p \to \infty$, the random variables $V_{i_1 i_2}$ are asymptotically jointly normally distributed with zero means and equal variances given by

$$\begin{aligned} \sigma_p^2 &= 2p(\lambda_4 + 3\lambda_2^2) + 8\lambda_2\|\mu\|^2 - 4\lambda_3 \sum_{j=1}^{p} \mu_j \\ &= 6p\lambda_2^2 + 6\lambda_2\|\mu\|^2 + 2\sum_{j=1}^{p} E\{(\delta - \mu_j)^2\delta^2\} \asymp \max(p, \|\mu\|^2), \end{aligned} \tag{3.6}$$

where $\lambda_k = E(\delta^k)$.

Property (3.5) is a standard negligibility condition in central limit theory. It holds quite generally, e.g. if c can be chosen so large that $\max_{1 \leqslant j \leqslant p} |\mu_j| = O(p^{1/2-\eta})$ where $\eta \geqslant 1/2c$.

The asymptotic relation $\sigma_p^2 \asymp \max(p, \|\mu\|^2)$ in expression (3.6) is defined to mean that the ratio of the left- and right-hand sides is bounded away from 0 and ∞ as p increases, for arbitrary choice of the quantities μ_j, with the latter potentially depending on p. The correctness of this asymptotic relation is a consequence of the second of the two identities for σ_p^2 in expression (3.6).

These properties immediately imply limit results for the probabilities π_{Y1} and π_{Y2}, defined at expression (2.2). In particular,

$$\begin{aligned} \pi_{Y1} &= P_Y(N_{i_1 i_2} \leqslant -\sigma_p^{-1}\|\mu\|^2 \text{ for all } 1 \leqslant i_1 \leqslant m \text{ and } 1 \leqslant i_2 \leqslant n) + o(1), \\ \pi_{Y2} &= P_Y(N_{i_1 i_2} \leqslant -\sigma_p^{-1}\|\mu\|^2 \text{ for some } 1 \leqslant i_1 \leqslant m \text{ and } 1 \leqslant i_2 \leqslant n) + o(1), \end{aligned} \tag{3.7}$$

as $p \to \infty$ for fixed m and n, where the variables $N_{i_1 i_2}$ are jointly normal $N(0, 1)$. When using expression (3.7), e.g. the first formula there, it is helpful to observe that, conditional on Z, $\|X_{i_1} - Z\| - \|Y_{i_2} - Z\|^2$ equals the difference between two independent random variables. Therefore, noting equation (3.3), we can write

$$P_Y(N_{i_1, i_2} \leqslant -\sigma_p^{-1}\|\mu\|^2 \text{ for all } i_1, i_2) = E[P\{N_1(i_1) \leqslant N_2(i_2) - \sigma_p^{-1}\|\mu\|^2 \text{ for all } i_1, i_2 \,|\mathcal{F}\}],$$

where, conditional on the σ-field $\mathcal{F}$, $N_1(i_1)$ and $N_2(i_2)$ are non-degenerate, independent and normally distributed variables and, unconditionally, $N_1(i_1) - N_2(i_2)$ is normal $N(0, 1)$. (In fact, these results continue to hold if we replace $\mathcal{F}$ by the trivial σ-field.)

It follows from the asymptotic relation in expression (3.6) that $\sigma_p^{-1}\|\mu\|^2 \to \infty$ if and only if $p = o(\|\mu\|^4)$. Hence, in view of expression (3.7), for all classifiers satisfying condition (2.1),

$$P_Y\{\mathcal{C}(Z) = \Pi_X\} \to 0 \qquad \text{if and only if } p = o(\|\mu\|^4) \text{ as } p \to \infty.$$

This result, and its analogue when Z is drawn from Π_X, rather than from Π_Y, imply property (3.2).

3.3. Properties of the threshold-based classifier $\mathcal{C}^{\rm tr}$

It is possible to develop theory describing $\mathcal{C}^{\rm tr}$ in the case where non-zero μ_js take a range of values. If, for all p, the number of possible, distinct non-zero values equals a fixed integer $k \geqslant 1$, then modified versions of the results in Sections 3 and 4 can be deduced without difficulty. The case where the number of 'subsignals' is unboundedly large is more complex, however. For simplicity we shall take $k = 1$, and assume that

$\mu_j = \nu > 0$ for q distinct indices j, and $\mu_j = 0$ for the remaining $p - q$ indices, where

(a) $\nu \geqslant t$,
(b) $t = t(p) \to \infty$ as p increases,
(c) $q = q(p)$ satisfies $q \to \infty$ and $1 \leqslant q \leqslant cp$, with $0 < c < 1$ fixed, and
(d) the distribution of δ is unbounded to the right. (3.8)

Part (a) of assumption (3.8) asks that the threshold t not exceed the size of the effect ν that causes the populations Π_X and Π_Y to differ. In practice, thresholds are chosen conservatively, ensuring that this constraint holds. Taking $t > \nu$ would not reflect practical motivation. Part (b) of assumption (3.8) serves only to make the problem different from the non-thresholded form; if t were taken to be fixed then first-order asymptotic properties of $\mathcal{C}^{\rm tr}$ would be equivalent to those of $\mathcal{C}$, discussed in Section 3.2. Part (c) asserts that the number of indices j for which $\mu_j \neq 0$ is not fixed, but is nevertheless a relatively small fraction of the total. Part (d) makes the classification problem non-degenerate, by preventing thresholding at t from removing, with probability 1, all vector components that correspond to an index j for which $\mu_j = 0$.

Assuming result (3.8) and a negligibility condition (3.12), below, we shall shortly prove that the following thresholded data version of property (3.2) holds:

> the probability that the classifier $\mathcal{C}^{\rm tr}$ correctly classifies a new data value from Π_X or Π_Y converges to 1 if and only if $p = o(\tau)$ as $p \to \infty$, (3.9)

where

$$\tau = (q\nu^2)^2 / E\{\delta^4\, I(\delta > t)\}. \tag{3.10}$$

Again this holds for general thresholded data, distance-based classifiers, e.g. those based on the support vector machine, distance-weighted discrimination or kth-nearest-neighbour classifiers.

Once more the assumption, in Section 3.1, that the disturbances δ_{ij}, ε_{ij} and δ_j all have the same distribution can be relaxed. In particular, property (3.9) remains true if the respective distribution functions F_{ij}, G_{ij} and F_j have the property that $1 - F_{ij}$, $1 - G_{ij}$ and $1 - F_j$ are uniformly bounded above and below by constant multiples of $1 - H$, where H is a fixed distribution for which condition (3.12) holds.

To derive property (3.9), note that the thresholded data analogue of the representation of distances at equation (3.3) is

$$\|X_{i_1}^{\rm tr} - Z^{\rm tr}\|^2 - \|Y_{i_2}^{\rm tr} - Z^{\rm tr}\|^2 = V_{i_1 i_2}^{\rm tr} + \|\mu^{\rm tr}\|^2, \tag{3.11}$$

where $\mu^{\rm tr} = (\mu_j^{\rm tr})$ is a p-vector, $\mu_j^{\rm tr} = E\{\eta_j\, I(\eta_j > t)\} - E\{\delta\, I(\delta > t)\}$, $\eta_j = \mu_j + \delta_j$ and the random variable $V_{i_1 i_2}^{\rm tr}$ has zero mean. Concise formulae for $V_{i_1 i_2}^{\rm tr}$ and its variance $(\sigma_p^{\rm tr})^2$ will be given in Section 5.2.

In the context of assumption (3.8), and assuming, in place of condition (3.5), that for some $c \geqslant 2$ we have $E(\delta^{4c}) < \infty$ and

$$[p\, E\{\delta^4\, I(\delta > t)\} + q\nu^4]^{-c}\, p\, E\{\delta^{4c}\, I(\delta > t)\} \to 0, \tag{3.12}$$

it can be proved that the variables $V_{i_1 i_2}^{\mathrm{tr}}$ are asymptotically jointly normally distributed with zero means and equal variances satisfying

$$(\sigma_p^{\mathrm{tr}})^2 \asymp p\, E\{\delta^4\, I(\delta > t)\} + q\nu^4. \tag{3.13}$$

See Section 5.2 for details. Property (3.12) holds quite generally, e.g. if c can be chosen so large that $p = O(q^c)$.

In view of condition (3.13), condition (3.7) holds for the new classifier $\mathcal{C}^{\mathrm{tr}}$, provided that we replace $\sigma_p^{-1}\|\mu\|^2$ there by

$$(\sigma_p^{\mathrm{tr}})^{-1}\|\mu^{\mathrm{tr}}\|^2 \asymp \left[\frac{q^2\nu^4}{p\, E\{\delta^4\, I(\delta > t)\} + q\nu^4}\right]^{1/2},$$

which diverges to ∞ if and only if $p = o(\tau)$, where τ is given by equation (3.10). This implies property (3.9).

4. Comparison of classifiers

4.1. Theoretical comparison of $\mathcal{C}$, $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$

Here we show that results (3.2) and (3.9), and their analogue for classifiers that are based on higher criticism, provide concise classification boundaries describing the relative performances of the classifiers $\mathcal{C}$, $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$. For definiteness we shall confine attention to the setting where assumption (3.8) holds, implying that the means μ_j are either 0 or ν for $p - q$ and q components respectively. We shall assume that the distributions of X_{ij} and Y_{ij} are $\mathrm{GN}_\gamma(0, 1)$ and $\mathrm{GN}_\gamma(\mu_j, 1)$ respectively.

The case of heavy-tailed distributions is similar. For example, if the common distribution function F of $X - EX$ and $Y - EY$ satisfies $1 - F(x) = Cx^{-\alpha}$ for constants $C > 0$ and $\alpha > 2$ and all sufficiently large x, then it can be shown that the classification boundaries for $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$ are identical, just as they are in the Subbotin case when $\gamma \leqslant 1$. This result can be generalized to a much larger class of heavy-tailed distributions and suggests that thresholded nearest neighbour methods are difficult to beat in cases where the marginal distributions are not particularly light tailed.

Specifically, assume that $q \sim \text{constant} \times p^{1-\beta}$ where $\frac{1}{2} < \beta < 1$, and take the threshold to be $t = \{\gamma r\, \log(p)\}^{1/\gamma}$ and the mean to be $\nu = \{\gamma s\, \log(p)\}^{1/\gamma}$, where $0 < r < s \leqslant 1$ denote constants. The inequality $s \leqslant 1$ is imposed without loss of generality, since, for any $\gamma' > \gamma$, with probability converging to 1 as $p \to \infty$ the maximum absolute value of p $\mathrm{GN}_\gamma(0, 1)$ random variables is strictly less than $\{\gamma'\, \log(p)\}^{1/\gamma}$.

The following result describes the relative performances of $\mathcal{C}$, $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$. Note that (a) and (b) below hold for all fixed values of m and n, whereas (c) requires $m = n = 1$:

necessary and sufficient conditions for the classifiers

(a) $\mathcal{C}$,
(b) $\mathcal{C}^{\mathrm{tr}}$ and
(c) $\mathcal{C}^{\mathrm{hc}}$

to produce asymptotically correct results are

(a) $1 - 2\beta > 0$,
(b) $1 - 2\beta + s > 0$ and, when $m = n = 1$,

(c)
(i) for $0 < \gamma \leqslant 1$, $1 - 2\beta + s > 0$, and
(ii) for $\gamma > 1$, $(1 - 2\beta)(2^{1/(\gamma-1)} - 1)^{\gamma-1} + 2s \geqslant 0$

when $\frac{1}{2} < \beta \leqslant 1 - 2^{-\gamma/(\gamma-1)}$, and $s > \{1 - (1-\beta)^{1/\gamma}\}^{\gamma}$ when $1 - 2^{-\gamma/(\gamma-1)} \leqslant \beta < 1$ respectively. (4.1)

The results in property (4.1) are necessary and sufficient in the sense that they hold under the strict inequalities given, and fail if those inequalities are strictly reversed. A derivation of property (4.1) will be given in Section 5.3.

4.2. Discussion of property (4.1)

Note that the boundaries in (a)–(c) above become steadily lower as we move through the sequence $\mathcal{C}$, $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$, implying that the respective classifiers are successively more sensitive, and in particular can produce asymptotically correct classification for successively smaller values of q, for a given value of s. When $0 < \gamma \leqslant 1$ and $m = n = 1$ the classification boundary for the higher criticism classifier is identical to that for the thresholded, distance-based method.

In the case $\gamma = 2$, corresponding to normal data, Fig. 1 graphs the classification boundaries corresponding to parts (b) and (c) of property (4.1), i.e. to the classifiers $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$ (the latter

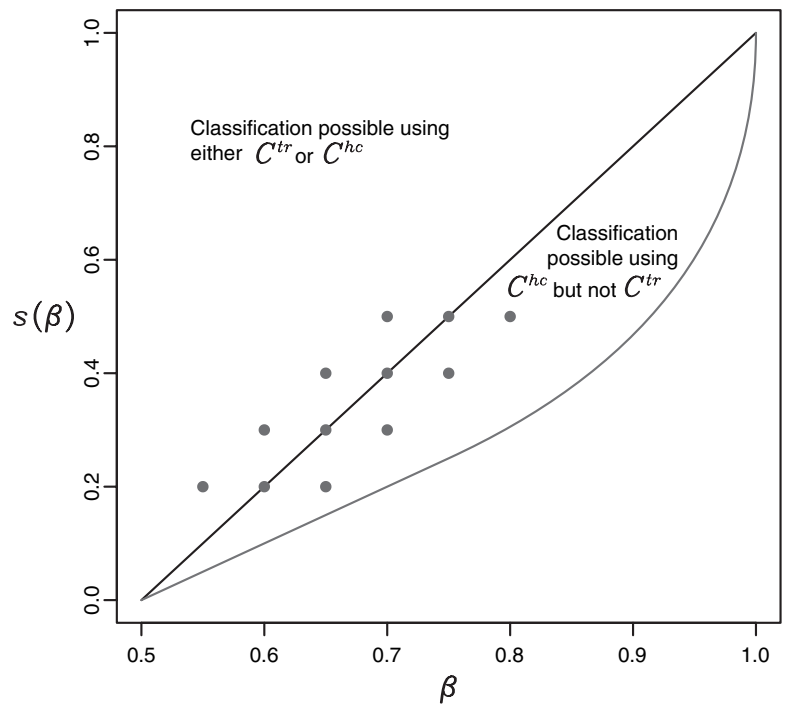

Fig. 1. Classification boundaries for $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$ when $\gamma = 2$: the upper line is a graph of the linear function $s = s(\beta)$ defined by $1 - 2\beta + s = 0$, which is the classification boundary for $\mathcal{C}^{\mathrm{tr}}$; the lower curve is a graph of the function $s = s(\beta)$ given by $1 - 2\beta + 2s \geqslant 0$, for $\frac{1}{2} < \beta < \frac{3}{4}$, and $s \geqslant \{1 - \sqrt{(1-\beta)}\}^2$, for $\frac{3}{4} < \beta < 1$, this being the classification boundary for $\mathcal{C}^{\mathrm{hc}}$ in the case of $N(0,1)$ data (i.e. $\mathrm{GN}_{\gamma}(0,1)$ data when $\gamma = 2$); inscriptions such as 'classification possible' indicate that 'for pairs (β, s) in this region, the probability of correct classification converges to 1 as p diverges' (•, positions of the point pairs (β, s) that are used in the numerical work in Section 4.4)

when $m = n = 1$). The domain $\frac{1}{2} < \beta < 1$ that is represented on the horizontal axis of Fig. 1 expresses β-values that are immediately to the right of the range $0 < \beta < \frac{1}{2}$, where estimation, and classification by the standard classifier $\mathcal{C}$ are both possible.

The two classification boundaries in Fig. 1 meet at $\beta = \frac{1}{2}$ and at $\beta = 1$. This feature expresses the fact that, at either extremity of the region corresponding to 'classifiable but not estimable differences', higher criticism classifiers perform similarly to their threshold-based counterparts. The lower curve in Fig. 1 is exactly that in either panel of Fig. 1 of Donoho and Jin (2004). Versions of Fig. 1 for general $\gamma > 1$ are similar, with of course the diagonal line, depicting the classification boundary for $\mathcal{C}^{\rm tr}$, in the same place. As $\gamma \downarrow 1$ the lower curve increases to the diagonal line.

4.3. Formulation of property (4.1) for empirical choice of t

For simplicity, in the discussion above we have assumed that an appropriate threshold t is given. That assumption is readily removed, however. We shall suggest an approach to selecting t empirically, which is valid under the models that were discussed in Section 4.1 and when at least one of m and n exceeds 1.

Assuming that $m = 2$ and $n = 1$, and using notation from Section 2.2, fix $\xi \in (0, 1)$ and define $\hat{t}$ to equal the infimum of values of t for which

$$\|X_1^{\rm tr} - X_2^{\rm tr}\| \exp\{\log(p)^{\xi}\} \leqslant \|Y_1^{\rm tr} - X_2^{\rm tr}\|, \tag{4.2}$$

using any default value, e.g. $\hat{t} = 0$, if no such t exists. Then, the classifiers $\mathcal{C}$, $\mathcal{C}^{\rm tr}$ and $\mathcal{C}^{\rm hc}$ give asymptotically correct classification under the respective conditions that are stated in property (4.1), and fail to give correct results if the inequalities there are strictly reversed. A proof of this result can be based on the fact that, for some $\eta \in (0, 1)$ and all $\zeta > 0$, and with probability converging to 1 as $n \to \infty$, $\hat{t}^{\gamma} > t_0^{\gamma} + \log(p)^{\eta}$ and $\hat{t}^{\gamma} < t_0^{\gamma} + \zeta \log(p)$, where $t_0 = \{\gamma(2\beta - 1) \log(p)\}^{1/\gamma}$.

If, rather than $m = 2$ and $n = 1$, we have $m = 1$ and $n = 2$, then the roles of X and Y in inequality (4.2) should be reversed. When m and n take larger values than these, the additional information can be incorporated by making appropriate modifications to inequality (4.2). Alternative approaches can deal with the case where $m = n = 1$; they are elementary if the classifier is applied not to the thresholded components $X_{ij}^{\rm tr}$ and $Y_{ij}^{\rm tr}$ but to the indicator functions, $I(X_{ij} > t)$ and $I(Y_{ij} > t)$, for which result (4.1) continues to apply.

4.4. Numerical comparison of $\mathcal{C}^{\rm tr}$ and $\mathcal{C}^{\rm hc}$

It is straightforward to show numerically that $\mathcal{C}^{\rm tr}$ can outperform $\mathcal{C}$. However, since the classification boundaries for $\mathcal{C}^{\rm tr}$ and $\mathcal{C}^{\rm hc}$ are close together then it is more challenging to show that the results that were described in Section 4.1 are reflected in numerical properties of $\mathcal{C}^{\rm tr}$ and $\mathcal{C}^{\rm hc}$. We shall do this in the setting that was explored by Donoho and Jin (2004), who took $p = 10^6$ (Donoho and Jin's notation replaces our p by their n) and $\gamma = 2$.

To construct Table 1 we chose pairs (β, s) that lay close to, and either above or below, the diagonal line in Fig. 1. This is the region where one or both of $\mathcal{C}^{\rm tr}$ and $\mathcal{C}^{\rm hc}$ can be expected to experience difficulty. The pairs fell into three classes:

(a) $(\beta, s) = (0.55, 0.2)$, (0.6, 0.3), (0.65, 0.4), (0.7, 0.5), each distant 0.1 above the diagonal line and, in this order, moving steadily to the right in Fig. 1, in which direction we expect classifiers to experience increasing difficulty;

(b) $(\beta, s) = (0.6, 0.2), (0.65, 0.3), (0.7, 0.4), (0.75, 0.5)$, all of these points lying on the diagonal line, where classification should be marginal for $\mathcal{C}^{\mathrm{tr}}$ but feasible for $\mathcal{C}^{\mathrm{hc}}$, but becoming increasingly difficult as we move through the points in the given order;

(c) $(\beta, s) = (0.65, 0.2), (0.7, 0.3), (0.75, 0.4), (0.8, 0.5)$, all of them 0.1 below the diagonal line.

For the classifier $\mathcal{C}^{\mathrm{tr}}$ we used $t = \{2r \log(p)\}^{1/2}$ where $r = 0.1$ when $s = 0.2$, 0.3, 0.4, and $r = 0.2$ when $s = 0.5$. These choices are not optimal, in the sense that they do not minimize classification error. However, the error rates are not far from the minimal rates, and the choices of r, being no more than half the value of s, are indicative of thresholds that might be used in practice.

We took $m = n = 10$ throughout and considered the performance of the classifiers $\mathcal{C}^{\mathrm{tr}}$ and $\mathcal{C}^{\mathrm{hc}}$ when the new value of Z came from Π_Y. In this setting, Table 1 gives the number of correct classifications for 100 independent realizations of Z in the respective cases, each value being the result of averaging results from 30 simulations, with standard errors given in parentheses.

As expected, the classification results in classes (a) and (b) are quite good, especially for $\mathcal{C}^{\mathrm{hc}}$, although they show deterioration as β is increased. The latter trend for $\mathcal{C}^{\mathrm{hc}}$ is also observed in cases (b) and (c). Results in case (c) are better than expected for $\mathcal{C}^{\mathrm{tr}}$, which has success rates of between 58% and 73% for the three pairs (β, s) for which, asymptotically, it does no better than 50%. Of course, in this setting $\mathcal{C}^{\mathrm{hc}}$ performs better still, with success rates between 68% and 75%. The percentages of correct results in case (b) lie between those in cases (a) and (c), for given values of β. The standard errors reveal that results for the classifier $\mathcal{C}^{\mathrm{hc}}$ are markedly less variable than those for $\mathcal{C}^{\mathrm{tr}}$.

In applying $\mathcal{C}^{\mathrm{hc}}$ to real data on genomic differential expression, we found that the classifier was generally outperformed by thresholded nearest neighbour methods. This may have been due to failure of the assumption of normality; the data were relatively heavy tailed. In practice, biologists tend to threshold such data, using their experience to determine the size of the threshold. In this connection it should be noted that, in problems where the presence of the signal can be detected but not estimated consistently, it is, almost by definition, difficult to select the threshold empirically—the threshold must be chosen strictly less than the unknowable value of the signal. From these viewpoints the work in the present paper provides theoretical under-

Table 1. Means and standard errors (in parentheses) of classification rates (percentage correctly predicted) using $\mathcal{C}^{\mathrm{hc}}$ and a nearest neighbour version of $\mathcal{C}^{\mathrm{tr}}$ for various pairs (β, s)

s	*Results for the following values of* β*:*											
	$\beta = 0.55$		$\beta = 0.60$		$\beta = 0.65$		$\beta = 0.70$		$\beta = 0.75$		$\beta = 0.80$	
	$\mathcal{C}^{\mathrm{hc}}$	$\mathcal{C}^{\mathrm{tr}}$	$\mathcal{C}^{\mathrm{hc}}$	$\mathcal{C}^{\mathrm{tr}}$	$\mathcal{C}^{\mathrm{hc}}$	$\mathcal{C}^{\mathrm{tr}}$	$\mathcal{C}^{\mathrm{hc}}$	$\mathcal{C}^{\mathrm{tr}}$	$\mathcal{C}^{\mathrm{hc}}$	$\mathcal{C}^{\mathrm{tr}}$	$\mathcal{C}^{\mathrm{hc}}$	$\mathcal{C}^{\mathrm{tr}}$
0.2	98.6 (0.2)	96.0 (1.3)	84.8 (0.7)	77.7 (4.3)	68.1 (0.8)	68.9 (5.9)						
0.3			99.7 (0.1)	96.3 (1.2)	90.3 (0.5)	78.4 (4.9)	73.9 (0.9)	73.2 (5.7)				
0.4					99.7 (0.1)	86.1 (3.1)	92.7 (0.5)	77.8 (4.7)	74.9 (1.0)	58.2 (7.0)		
0.5							99.4 (0.1)	80.0 (6.5)	89.4 (0.7)	63.8 (7.1)	73.5 (1.0)	71.0 (6.6)

pinning for using thresholded, distance-based classifiers, such as nearest neighbour methods, and for employing experience rather than empirical methods to select the threshold.

5. Technical arguments

5.1. Properties of $V_{i_1 i_2}$, defined at equation (3.4)

Derivation of the variance formulae at expression (3.6) is straightforward. Using the fact that $V_{i_1 i_2}$ is a sum of independent random variables with zero mean, asymptotic normality can be proved from Lindeberg's theorem. Since fourth moments are finite then, to establish Lindeberg's condition, it is necessary only to show that

$$\sum_{j=1}^{p} E\{|\delta\mu_j/\sigma_p|^2\, I(|\delta\mu_j/\sigma_p| > 1)\} \to 0,$$

for which condition (3.5) is sufficient. Here we have used the property $\sigma_p^2 \asymp \max(p, \|\mu\|^2)$, which is taken from expression (3.6).

5.2. Properties of $V^{\mathrm{tr}}_{i_1 i_2}$, defined at equation (3.11)

The quantity $V^{\mathrm{tr}}_{i_1 i_2}$ is given by

$$V^{\mathrm{tr}}_{i_1 i_2} = \sum_{j=1}^{p} \{(\delta^{\mathrm{tr}}_{i_1 j})^2 - (\eta^{\mathrm{tr}}_{i_2 j})^2 - 2\delta^{\mathrm{tr}}_{i_1 j}\mu^{\mathrm{tr}}_j - 2(\delta^{\mathrm{tr}}_{i_1 j} - \mu^{\mathrm{tr}}_j - \eta^{\mathrm{tr}}_{i_2 j})\eta^{\mathrm{tr}}_j\},$$

where $\delta^{\mathrm{tr}}_{ij} = (1-E)\delta_{ij}\, I(\delta_{ij} > t)$, $\eta^{\mathrm{tr}}_{ij} = (1-E)\eta_{ij}\, I(\eta_{ij} > t)$, $\eta^{\mathrm{tr}}_j = (1-E)\eta_j\, I(\eta_j > t)$, $\eta_{ij} = \mu_j + \varepsilon_{ij}$, $\eta_j = \mu_j + \delta_j$ and E denotes the expectation operator. Writing δ^{tr} for $\delta^{\mathrm{tr}}_{11}$, the variance of $\mathrm{var}(V^{\mathrm{tr}}_{i_1 i_2})$ can be shown to equal

$$\begin{aligned}
(\sigma^{\mathrm{tr}}_p)^2 &= \sum_{j=1}^{p} [\mathrm{var}\{(\delta^{\mathrm{tr}})^2\} + \mathrm{var}\{(\eta^{\mathrm{tr}}_j)^2\} + 4\,\mathrm{var}(\delta^{\mathrm{tr}})(\mu^{\mathrm{tr}}_j)^2 + 4\{\mathrm{var}(\delta^{\mathrm{tr}}) + (\mu^{\mathrm{tr}}_j)^2 + \mathrm{var}(\eta^{\mathrm{tr}}_j)\}\,\mathrm{var}(\eta^{\mathrm{tr}}_j) \\
&\quad - 4\, E\{(\delta^{\mathrm{tr}})^3\}\mu^{\mathrm{tr}}_j] \\
&\asymp \sum_{j=1}^{p} (E\{\delta^4\, I(\delta > t)\} + \mu_j^4\, P(\delta > t - \mu_j) + E\{\delta^4\, I(\delta > t - \mu_j)\} + [E\{\delta^2\, I(\delta > t)\} \\
&\quad + \mu_j^2\, P(\delta > t - \mu_j) + E\{\delta^2\, I(\delta > t - \mu_j)\}]|\mu_j|^2) \\
&\asymp p\, E\{\delta^4\, I(\delta > t)\} + q\nu^4.
\end{aligned}$$

This establishes condition (3.13).

A similar argument shows that if moments of order $4c$ are finite then the version of Lindeberg's condition in the present setting is satisfied if $(\sigma^{\mathrm{tr}}_p)^{-2c} a_c \to 0$, where

$$\begin{aligned}
a_c &\asymp \sum_{j=1}^{p} (E\{\delta^{4c}\, I(\delta > t)\} + \mu_j^{4c}\, P(\delta > t - \mu_j) + E\{\delta^{4c}\, I(\delta > t - \mu_j)\} \\
&\quad + [E\{\delta^{2c}\, I(\delta > t)\} + |\mu_j|^{2c} + E\{\delta^{2c}\, I(\delta > t - \mu_j)\}]\,[|\mu_j|^{2c}\, P(\delta > t - \mu_j) \\
&\quad + E\{\delta^{2c}\, I(\delta > t - \mu_j)\}]) \\
&\asymp p\, E\{\delta^{4c}\, I(\delta > t)\} + q\nu^{4c}.
\end{aligned}$$

Now, $(\sigma^{\mathrm{tr}}_p)^{-2c}[p\, E\{\delta^{4c}\, I(\delta > t)\} + q\nu^{4c}] \to 0$ if and only if condition (3.12) holds, establishing the sufficiency of condition (3.12) for the central limit theorem.

5.3. Derivation of result (4.1)

Let $b(s, \beta) > 0$ denote the formula for a general detection boundary that is treated in result (4.1). In case (a), property (3.2) implies that asymptotically correct classification is possible if and only if $p = o(\|\mu\|^4)$, which is equivalent to $p = o\{p^{2(1-\beta)} \log(p)^{4/\gamma}\}$ and hence to $1 - 2\beta \geqslant 0$. In case (b), property (3.9) implies that asymptotically correct classification is possible if and only if $p = o\{(q^2/p^{-r}) \log(p)^c\}$, where c is a constant or, equivalently, if and only if $p = o\{p^{2(1-\beta)+r} \log(p)^c\}$; call this result (R). Since $r < s$ but can be chosen arbitrarily close to s, then result (R) can hold if $1 - 2\beta + s > 0$ but not if $1 - 2\beta + s < 0$.

Finally we give an outline proof of part (c) of result (4.1), the case of $\mathcal{C}^{\rm hc}$. Suppose that the model (3.1) obtains, and that Z is drawn from Π_Y, which is characterized by the fact that just q of the μ_js equal $\{\gamma s \log(p)\}^{1/\gamma}$, and each of the other $p - q$ μ_js equals 0. If $s > r$ and (β, s) lies strictly above the boundary $b(s, \beta) = 0$ that is given in (c) of result (4.1), then for some $\varepsilon > 0$ (chosen sufficiently small), and with probability converging to 1, $\mathrm{hc}_X < -p^\varepsilon$ for all sufficiently large p. However, for each $\varepsilon > 0$ and with probability converging to 1, $|\mathrm{hc}_Y| < p^\varepsilon$ for all sufficiently large p. Together the results imply that $\mathrm{hc}_Y > \mathrm{hc}_X$, with probability converging to 1 as $p \to \infty$. Equivalently, the probability that Z is correctly classified as coming from Π_Y converges to 1. Similarly it can be shown that if Z is drawn from Π_X then the probability that Z is classified as coming from Π_X converges to 1.

References

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

Baik, J., Ben Arous, G. and Peche, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**, 1643–1697.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, Jr, M. and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natn. Acad. Sci. USA*, **97**, 262–267.

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

Cayon, L., Banday, A. J., Jaffe, T., Eriksen, H. K., Hansen, F. K., Gorski, K. M. and Jin, J. (2006) No higher criticism of the Bianchi corrected WMAP data. *Mnthly Notes R. Astronom. Soc.*, **369**, 598–602.

Cayon, L., Jin, J. and Treaster, A. (2005) Higher criticism statistic: detecting and identifying non-Gaussianity in the WMAP first year data. *Mnthly Notes R. Astronom. Soc.*, **362**, 826–832.

Choi, S.-H. and Rockett, P. (2002) The training of neural classifiers with condensed datasets. *IEEE Trans. Syst. Man Cybernet.* B, **32**, 202–206.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.

Donoho, D. L. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

Efron, B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Statist. Ass.*, **99**, 96–104.

El Karoui, N. (2005) Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Phys. Polon.* B, **36**, 2681–2697.

Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B, **67**, 427–444.

Ibragimov, I. A. (1962) Some limit theorems for stationary processes. *Theory Probab. Appl.*, **4**, 349–382.

Ingster, Yu. I. (1999) Minimax detection of a signal for l^n-balls. *Math. Meth. Statist.*, **7**, 401–428.

Ingster, Yu. I. (2001) Adaptive detection of a signal of growing dimension: I. *Math. Meth. Statist.*, **10**, 395–421.

Ingster, Yu. I. (2002) Adaptive detection of a signal of growing dimension: II. *Math. Meth. Statist.*, **11**, 37–68.

Jin, J. (2002) Detection boundary for sparse mixtures. *Manuscript*. Unpublished.

Jin, J. (2005) Detecting a target in very noisy data from multiple looks. In *A Festschrift to Honor Herman Rubin* (ed. A. Dasgupta), pp. 255–286. Beachwood: Institute of Mathematical Statistics.

Jin, J. (2006) Higher criticism statistic: theory and applications in non-Gaussian detection. In *Proc. PHYSTAT 2005: Statistical Problems in Particle Physics, Astrophysics and Cosmology* (eds L. Lyons and M. K. Ünel). Singapore: World Scientific Publishing.

Jin, J., Starck, J.-L., Donoho, D. L., Anghanim, N. and Forni, O. (2005) Cosmological non-Gaussian signature detection: comparing performance of different statistical tests. *EURASIP J. Appl. Signal Process.*, **15**, 2470–2485.
Johnstone, I. M. (2001) On the distribution of the largest principal component. *Ann. Statist.*, **29**, 295–327.
Marron, J. S., Todd, M. and Ahn, J. (2005) Distance weighted discrimination. *Manuscript*. Unpublished.
Meinshausen, M. and Rice, J. (2004) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, **34**, 373–393.
Murtagh, F. (1985) *Multidimensional Clustering Algorithms*. Wurzburg: Physica.
Politis, D. N., Romano, J. P. and Wolf, M. (1997) Subsampling for heteroskedastic time series. *J. Econometr.*, **81**, 281–317.
Sarandasa, H. and Altan, S. (1998) The analysis of small-sample multivariate data. *J. Biopharm. Statist.*, **8**, 163–186.
Schölokopf, B. and Smola, A. (2001) *Learning with Kernels*. Cambridge: MIT Press.
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.