HIGHER CRITICISM IN THE CONTEXT OF UNKNOWN DISTRIBUTION, NON-INDEPENDENCE AND CLASSIFICATION

Aurore Delaigle [1,2] Peter Hall[2]

ABSTRACT. Higher criticism has been proposed as a tool for highly multiple hypothesis testing or signal detection, initially in cases where the distribution of a test statistic (or the noise in a signal) is known and the component tests are statistically independent. In this paper we explore the extent to which the assumptions of known distribution and independence can be relaxed, and we consider too the application of higher criticism to classification. It is shown that effective distribution approximations can be achieved by using a threshold approach; that is, by disregarding data components unless their significance level exceeds a sufficiently high value. This method exploits the good relative accuracy of approximations to light-tailed distributions. In particular, it can be effective when the true distribution is founded on something like a Studentised mean, or on an average of related type, which is commonly the case in practice. The issue of dependence among vector components is also shown not to be a serious difficulty in many circumstances.

KEYWORDS. Autoregression, distribution approximation, moving-average process, multiple comparison, signal detection, sparsity, Studentised ratios, thresholding, time-series.

SHORT TITLE. Higher criticism.

[1] Department of Mathematics, University of Bristol, Bristol, BS8 4JS, UK

[2] Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia

1 Introduction

Donoho and Jin (2004) developed higher-criticism methods for hypothesis testing and signal detection. Their methods are founded on the assumption that the test statistics are independent and, under the null hypothesis, have a known normal distribution. However, in some applications of higher criticism, for example to more elaborate hypothesis testing problems and to classification, these assumptions may not be tenable. For example, we may have to estimate the distributions from data, by pooling information from components that have "neighbouring" indices, and the assumption of independence may be violated.

Taken together, these difficulties place obstacles in the way of using higher-criticism methods for a variety of applications, even though those techniques have potential performance advantages. We describe the effects that distribution approximation and data dependence can have on results, and suggest ways of alleviating problems caused by distribution approximation. We show too that thresholding, where only deviations above a particular value are considered, can produce distinct performance gains. Thresholding permits the experimenter to exploit the greater relative accuracy of distribution approximations in the tails of a distribution, compared with accuracy towards the distribution's centre, and thereby to reduce the tendency of approximation errors to accumulate. Our theoretical arguments take sample size to be fixed and the number of dimensions, p, to be arbitrarily large.

Thresholding makes it possible to use rather crude distribution approximations. In particular, it permits the approximations to be based on relatively small sample sizes, either through pooling data from a small number of nearby indices, or by using normal approximations based on averages of relatively small datasets. Without thresholding, the distribution approximations used to construct higher-criticism signal detectors and classifiers would have to be virtually root-p consistent.

We shall provide theoretical underpinning for these ideas, and explore them numerically; and we shall demonstrate that higher criticism can accommodate significant amounts of local dependence, without being seriously impaired. We shall further show that, under quite general conditions, the higher-criticism statistic can

be decomposed into two parts, of which one is stochastic and of smaller order than p^ϵ for any positive ϵ, and the other is purely deterministic and admits a simple, explicit formula. This simplicity enables the effectiveness of higher criticism to be explored quite generally, for distributions where the distribution tails are heavy, and also for distributions that have relatively light tails, perhaps through being convolutions of heavy-tailed distributions. These comments apply to applications to both signal detection and classification.

In the contexts of independence and signal detection, Donoho and Jin (2004) used an approach alternative to that discussed above. They employed delicate, empirical-process methods to develop a careful approximation, on the $\sqrt{\log \log p}$ scale, to the null distribution of the higher-criticism statistic. It is unclear from their work whether the delicacy of the log-log approximation is essential, or whether significant latitude is available for computing critical points. We shall show that quite crude bounds can in fact be used, in both the dependent and independent cases. Indeed, any critical point on a scale that is of smaller order than p^ϵ, for each $\epsilon > 0$, is appropriate.

Higher-criticism methods for signal detection have their roots in unpublished work of Tukey; see Donoho and Jin (2004) for discussion. Optimal, but more tightly specialised, methods for signal detection were developed by Ingster (1999, 2001, 2002) and Ingster and Suslina (2003), broadly in the context of techniques for multiple comparison (see e.g. the methods of Bonferroni, Tukey (1953) and Scheffé (1959)), for simultaneous hypothesis testing (e.g. Efron (2004) and Lehmann *et al.* (2005)) and for moderating false-discovery rates (e.g. Benjamini and Hochberg (1995), Storey *et al.* (2005) and Abramovich *et al.* (2006)). Model-based approaches to the analysis of high-dimensional microarray data include those of Tseng *et al.* (2001), Huang *et al.* (2003), Fan *et al.* (2005b) and Fan and Fan (2007). Related work on higher criticism includes that of Meinshausen and Rice (2006) and Cai, Jin and Low (2007). Higher-criticism classification has been discussed by Hall *et al.* (2008), although this work assumed that test statistic distributions are known exactly. Applications of higher criticism to signal detection in astronomy include those of Jin *et al.* (2004), Cayón *et al.* (2005, 2006), Cruz *et al.* (2007)

and Jin (2006). Hall and Jin (2006) discussed properties of higher criticism under long-range dependence.

Our main theoretical results are as follows. Theorem 3.1, in section 3.1, gives conditions under which the higher-criticism statistic, based on a general approximation to the unknown test distributions, can be expressed in terms of its "ideal" form where the distributions are known, plus a negligible remainder. This result requires no assumptions about independence. Theorem 3.2, in section 3.2, gives conditions on the strength of dependence under which the higher-criticism statistic can be expressed as a purely deterministic quantity plus a negligible remainder. Theorem 3.3, in section 3.3, describes properties of the deterministic "main term" in the previous result. Discussion in sections 3.3 and 4 draws these three results together, and shows that they lead to a variety of properties of signal detectors and classifiers based on higher criticism. These properties are explored numerically in section 5.

2 Methodology

2.1 Higher-criticism signal detection

Assume we observe independent random variables $Z_1, \ldots, Z_p$, where each Z_j is normally distributed with mean μ_j and unit variance. We wish to test, or at least to assess the validity of, the null hypothesis H_0 that each μ_j equals ν_j, a known quantity, versus the alternative hypothesis that one or more of the μ_j are different from ν_j. If each ν_j equals zero then this context models signal detection problems where the null hypothesis states that the signal is comprised entirely of white noise, and the alternative hypothesis indicates that a nondegenerate signal is present.

A higher-criticism approach to signal detection and hypothesis testing, a two-sided version of a suggestion by Donoho and Jin (2004), can be based on the statistic

$$\mathrm{hc} = \inf_{u\,:\,\psi(u)>C} \psi(u)^{-1/2} \sum_{j=1}^{p} \{I(|Z_j - \nu_j| \le u) - \Psi(u)\}\,, \tag{2.1}$$

where $\Psi(u) = 2\,\Phi(u) - 1$ is the distribution function of $|Z_j - \nu_j|$ under H_0, Φ is the standard normal distribution function, $\psi(u) = p\,\Psi(u)\,\{1 - \Psi(u)\}$ equals the variance of $\sum_j \{I(|Z_j - \nu_j| \le u) - \Psi(u)\}$ under H_0, and C is a positive constant.

The statistic at (2.1) provides a way of assessing the statistical significance of p tests of significance. In particular, H_0 is rejected if hc takes too large a negative value. This test enjoys optimality properties, in that it is able to detect the presence of nonzero values of μ_j up to levels of sparsity and amplitude that are so high and so low, respectively, that no test can distinguish between the null and alternative hypotheses (Donoho and Jin, 2004).

2.2 Generalising and adapting to an unknown null distribution

When employed in the context of hypothesis testing (where the ν_js are not necessarily equal to zero), higher-criticism could be used in more general settings, where the centered Z_js are not identically distributed. Further, instead of assuming that the ν_js are prespecified, they could be taken equal to the jth component of the empirical mean of a set of n_W identically distributed random p-vectors $W_1, \ldots, W_{n_W}$, where $W_i = (W_{i1}, \ldots, W_{ip})$ has the distribution of $(Z_1, \ldots, Z_p)$ under the null hypothesis H_0. Here, H_0 asserts the equality of the mean components of the vector Z, and of the vectors $W_1, \ldots, W_{n_W}$, whose distribution is known except for the mean which is estimated by its empirical counterpart. There, we could redefine hc by replacing, in (2.1), ν_j by $\bar{W}_{\cdot j} = n_W^{-1} \sum_j W_{ij}$ and Ψ by the distribution Ψ_{Wj}, say, of $|Z_j - \bar{W}_{\cdot j}|$ under the null hypothesis. This gives, in place of hc at (2.1),

$$\mathrm{hc}_W = \inf_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \{I(|Z_j - \bar{W}_{\cdot j}| \le u) - \Psi_{Wj}(u)\}\,, \tag{2.2}$$

where

$$\psi_W(u) = \sum_{j=1}^{p} \Psi_{Wj}(u)\,\{1 - \Psi_{Wj}(u)\} \tag{2.3}$$

and, given $C, t > 0$, $\mathcal{U}_W = \mathcal{U}_W(C,t)$ is the set of u for which $u \ge t$ and $\psi_W(u) \ge C$. Here t denotes a threshold, and the fact that, in the definition of $\mathcal{U}_W$, we confine attention to $u > t$, means that we restrict ourselves to values of u for which distribution approximations are relatively accurate; see section 4.2 for details.

Further, in practical applications it is often unrealistic to argue that Ψ (respectively, Ψ_{Wj}), is known exactly, and we should replace Ψ in (2.1) and in ψ (respectively, Ψ_{Wj} in (2.2) and in ψ_W) by an estimator $\widehat{\Psi}$ of Ψ (respectively, $\widehat{\Psi}_{Wj}$

of Ψ_{Wj}). This leads to an empirical approximation, $\widehat{\mathrm{hc}} = \widehat{\mathrm{hc}}(C,t)$, to hc:

$$\widehat{\mathrm{hc}} = \inf_{u \in \widehat{\mathcal{U}}} \widehat{\psi}(u)^{-1/2} \sum_{j=1}^{p} \left\{ I(|Z_j - \nu_j| \le u) - \widehat{\Psi}(u) \right\},$$

where $\widehat{\psi}(u) = p\,\widehat{\Psi}(u)\,\{1 - \widehat{\Psi}(u)\}$ and $\widehat{\mathcal{U}} = \widehat{\mathcal{U}}(C,t)$ is the set of u for which $u \ge t$ and $\widehat{\psi}(u) \ge C$, and to an empirical approximation $\widehat{\mathrm{hc}}_W = \widehat{\mathrm{hc}}_W(C,t)$, to hc_W:

$$\widehat{\mathrm{hc}}_W = \inf_{u \in \widehat{\mathcal{U}}_W} \widehat{\psi}_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ I(|Z_j - \bar{W}_{\cdot j}| \le u) - \widehat{\Psi}_{Wj}(u) \right\}, \tag{2.4}$$

where

$$\widehat{\psi}_W(u) = \sum_{j=1}^{p} \widehat{\Psi}_{Wj}(u) \left\{ 1 - \widehat{\Psi}_{Wj}(u) \right\} \tag{2.5}$$

and $\widehat{\mathcal{U}}_W = \widehat{\mathcal{U}}_W(C,t)$ is the set of u for which $u \ge t$ and $\widehat{\psi}_W(u) \ge C$. Here t denotes a threshold, and the fact that, in the definition of $\mathcal{U}_W$, we confine attention to $u > t$, means that we restrict ourselves to values of u for which distribution approximations are relatively accurate; see section 4.2 for details.

Estimators of Ψ_{Wj} are, broadly speaking, of two types: either they depend strongly on the data, or they depend on the data only through the way these have been collected, for instance through sample size. In the first case, $\widehat{\Psi}_{Wj}$ might, for example, be computed directly from the data, for example by pooling vector components for nearby values of the index j. (In genomic examples, "nearby" does not necessarily mean close in terms of position on the chromosome; it is often more effectively defined in other ways, for example in the sense of gene pathways.)

Examples of the second type come from an important class of problems where the variables W_{ij} are obtained by averaging other data. For example, they can represent Studentised means, $W_{ij} = N_{Wj}^{1/2}\,\bar{U}_{Wij}/S_{Wij}$, or Student t statistics for two-sample tests, $W_{ij} = N_{Wj}^{1/2}\,(\bar{U}_{Wij,1} - \bar{U}_{Wij,2})/(S_{Wij,1}^2 + S_{Wij,2}^2)^{1/2}$, where, for $i = 1, \ldots, n_W$ and $j = 1, \ldots, p$, $\bar{U}_{Wij,k}$ and $S_{Wij,k}^2$, $k = 1, 2$ denote respectively the empirical mean and empirical variance of N_{Wj} independent and identically distributed data; or statistics computed in a related way. See e.g. Lönnstedt and Speed (2001), Storey and Tibshirani (2003), Fan *et al.* (2004) and Fan *et al.* (2005a).

In such cases, if Z_j and $W_{1j}, \ldots, W_{nj}$ were identically distributed, $Z_j - \bar{W}_{\cdot j}$ would be approximately normally distributed with variance $\tau_W = 1 + n_W^{-1}$, and $\widehat{\Psi}_{Wj}$ would be the distribution function of the normal $\mathrm{N}(0, \tau_W)$ distribution, not depending on the index j, and depending on the data only through the number n_W of observations. See section 3.3 for theoretical properties for this type of data.

Formula (2.5), giving an empirical approximation to the variance of the series on the right-hand side of (2.4), might seem to suggest that, despite the increased generality we are capturing by using empirical approximations to the distribution functions Ψ_{Wj}, we are continuing to assume that the vector components $W_{i1}, \ldots, W_{ip}$ are independent. However, independence is not essential. By choosing the threshold, t, introduced earlier in this section, to be sufficiently large, the independence assumption can be removed while retaining the validity of the variance approximation at (2.5). See section 4.2.

2.3 Classifiers based on higher criticism

The generality of the higher-criticism approximation in section 2.2 leads directly to higher-criticism methods for classification. To define the classification problem, assume we have data, in the form of independent random samples of p-dimensional vectors $\mathcal{X} = \{X_1, \ldots, X_{n_X}\}$ from population Π_X, and $\mathcal{Y} = \{Y_1, \ldots, Y_{n_Y}\}$ from population Π_Y, and a new, independent observation, Z, from either Π_X or Π_Y. (In our theoretical work the sample sizes n_X and n_Y will be kept fixed as p increases.) We wish to assign Z to one of the populations. In the conventional case where p is small relative to sample size, many different techniques have been developed for solving this problem. However, in the setting in which we are interested, where p is large and the sample size is small, these methods can be ineffective, and better classification algorithms can be obtained by using methods particularly adapted to the detection of sparse signals.

Let X_{ij}, Y_{ij} and Z_j denote the jth components of X_i, Y_i and Z, respectively. Assume that $\mu_{Xj} = E(X_{ij})$ and $\mu_{Yj} = E(Y_{ij})$ do not depend on i, that the distributions of the components are absolutely continuous, and that the distributions of the vectors $(X_{i_1 1} - \mu_{X1}, \ldots, X_{i_1 p} - \mu_{Xp})$, $(Y_{i_2 1} - \mu_{Y1}, \ldots, Y_{i_2 p} - \mu_{Yp})$

and $(Z_1 - E(Z_1), \ldots, Z_p - E(Z_p))$ are identical to one another, for $1 \le i_1 \le n_X$ and $1 \le i_2 \le n_Y$.

In particular, for each i_1, i_2 and j the distributions of $X_{i_1 j}$ and $Y_{i_2 j}$ differ only in location. This assumption serves to motivate methodology, and is a convenient platform for theoretical arguments. Of course, many other settings can be addressed, but they are arguably best treated using their intrinsic features. Instances of particular interest include those where each component distribution is similar to a Studentised mean. A particular representation of this type, involving only location changes, will be discussed extensively in section 3.3. Other variants, where non-zero location also entails changes in shape, can be treated using similar arguments, provided the shape-changes can be parametrised.

With W denoting X or Y we shall write V_{Wj} for a random variable having the distribution of $Z_j - \bar{W}_{\cdot j}$, given that Z is drawn from Π_W. If $n_X = n_Y$ then the distribution of V_{Wj} depends only on j, not on choice of W. Let $\bar{X}_{\cdot j} = {n_X}^{-1} \sum_i X_{ij}$ and define $\bar{Y}_{\cdot j}$ analogously. Let Ψ_{Wj} be the distribution function of $|V_{Wj}|$, and put $\Delta_{Wj}(u) = I(|Z_j - \bar{W}_{\cdot j}| \le u) - \Psi_{Wj}(u)$.

If Z is from Π_W then, for each j, $|Z_j - \bar{W}_{\cdot j}|$ has distribution function Ψ_{Wj}, and so, for each fixed u, $\Delta_{Wj}(u)$ has expected value zero. On the other hand, since the distributions of X_{ij} and Y_{ij} may differ in location, then, if Z is not from Π_W, $P(|Z_j - \bar{W}_{\cdot j}| \le u)$ may take a lesser value than it does when Z is from Π_W, with the result that the expected value of $\Delta_{Wj}(u)$ can be strictly negative. Provided an estimator of Ψ_{Wj} is available for $W = X$ and $W = Y$, this property can be used to motivate a classifier. In particular, defining $\widehat{\mathrm{hc}}_X$ and $\widehat{\mathrm{hc}}_Y$ as at (2.4), we should classify Z as coming from Π_X if $\widehat{\mathrm{hc}}_X \ge \widehat{\mathrm{hc}}_Y$, and as coming from Π_Y otherwise.

3 Theoretical properties

3.1 Effectiveness of approximation to hc_W by $\widehat{\mathrm{hc}}_W$

We start by studying the effectiveness of the approximation by $\widehat{\mathrm{hc}}_W$ to hc_W, where hc_W and $\widehat{\mathrm{hc}}_W$ are defined as at (2.2) and (2.4), respectively (arguments similar to those given here can be used to demonstrate the effectiveness of the approximation

by $\widehat{\text{hc}}$ to hc). To embed the case of hypothesis testing in that of classification, we express the problem of hypothesis testing as one where the vector Z comes from a population Π_Z, equal to either Π_X or Π_W, and where the data vectors $W_1, \ldots, W_{n_W}$ come from Π_W, with $\Pi_W = \Pi_Z$ under H_0, $\Pi_W = \Pi_T$ otherwise, and (Π_Z, Π_T) denoting one of (Π_X, Π_Y) or (Π_Y, Π_X). We assume throughout section 3 that each $\widehat{\Psi}_{Wj}$ is, with probability 1, a continuous distribution function satisfying $\widehat{\Psi}_{Wj}(x) \to 0$ as $x \downarrow 0$ and $\widehat{\Psi}_{Wj}(x) \to 1$ as $x \uparrow \infty$. We also make the following additional assumptions.

Condition A

(A1) The threshold, $t = t(p)$, varies with p in such a manner that: For each $C > 0$ and for $W = X$ and Y, $\sup_{u \in \mathcal{U}_W(C,t)} \psi_W(u)^{-1/2} \sum_j |\widehat{\Psi}_{Wj}(u) - \Psi_{Wj}(u)| = o_p(1)$.
(A2) For a constant $u_0 > 0$, for each of the two choices of W, and for all sufficiently large p, ψ_W is nonincreasing and strictly positive on $[u_0, \infty)$; and the probability that $\widehat{\psi}_W$ is nonincreasing on $[u_0, \infty)$ converges to 1 as $p \to \infty$.

The reasonableness of (A1) is taken up in section 4.2, below. The first part of (A2) says merely that ψ_W inherits the monotonicity properties of its component parts, $\Psi_{Wj}\,(1 - \Psi_{Wj})$. Indeed, if Ψ_{Wj} is the distribution function of a distribution that has unbounded support, then $\Psi_{Wj}\,(1 - \Psi_{Wj})$ is nonincreasing and strictly positive on $[u_0, \infty)$ for some $u_0 > 0$, and (A2) asks that the same be true of $\psi_W = \sum_j \Psi_{Wj}\,(1 - \Psi_{Wj})$. This is of course trivial if the distributions Ψ_{Wj} are identical. The second part of (A2) states that the same is true of the estimator, $\widehat{\psi}_W$, of ψ_W, which condition is satisfied if, for example, the observations represent Studentised means.

The next theorem shows that, under sufficient conditions, $\widehat{\text{hc}}_W$ is an effective approximation to hc_W. Note that we make no assumptions about independence of vector components, or about the population from which Z comes. In particular, the theorem is valid for data drawn under both the null and alternative hypotheses.

Theorem 3.1. *Let $0 < C_1 < C_2 < C_3 < \infty$ and $0 < C < C_3$, and assume that $\psi_W(t) \geq C_3$ for all sufficiently large p. If* (A1) *and* (A2) *hold then, with $W = X$*

or Y,

$$\widehat{\mathrm{hc}}_W(C,t) = \{1+o_p(1)\} \inf_{u \in \widehat{\mathcal{U}}_W(C,t)} \psi_W(u)^{-1/2} \times \sum_{j=1}^{p} \bigl\{ I(|Z_j - \bar{W}_{\cdot j}| \le u) - \Psi_{Wj}(u) \bigr\} + o_p(1)\,, \quad (3.1)$$

$$\begin{aligned} \mathrm{hc}_W(C_1,t) + o_p(1) &\le \{1+o_p(1)\}\, \widehat{\mathrm{hc}}_W(C_2,t) + o_p(1) \\ &\le \{1+o_p(1)\}\, \mathrm{hc}_W(C_3,t) + o_p(1)\,. \end{aligned} \quad (3.2)$$

We shall see in the next section that, in many cases of interest, when Z is not drawn from Π_W, the higher-criticism statistic hc_W tends, with high probability, to be negative and does not converge to zero as $p \to \infty$. Our results in the next section also imply that, when Z comes from Π_W, the last, added remainders $o_p(1)$ on the far right-hand sides of (3.1) and (3.2) are of smaller order than the earlier quantities on the right. Together, these properties justify approximating hc_W by $\widehat{\mathrm{hc}}_W$.

3.2 Removing the assumption of independence

We now study the properties of higher-criticism statistics in cases where the components are not independent. To illustrate the type of dependence that we have in mind, let us consider the case where Z is drawn from Π_W, and the variables $V_j = Z_j - \bar{W}_{\cdot j}$ form a mixing sequence with exponentially rapidly decreasing mixing coefficients. The case where the mixing coefficients decrease only polynomially fast, as functions of p, can also be treated; see Remark 3.3.

To give a specific example, note that the cases of moving-average processes or autoregressions, of arbitrary (including infinite) order, fall naturally into the setting of exponentially fast mixing. Indeed, assume for simplicity that the variables V_j form a stochastic process, not necessarily stationary, that is representable as

$$V_j = \sum_{k=1}^{\infty} \alpha_{jk}\, \xi_{j-k}\,,$$

where the α_{jk}'s are constants satisfying $|\alpha_{jk}| \le \text{const.}\, \rho^k$ for all j and k, $0 < \rho < 1$, and the disturbances ξ_j are independent with zero means and uniformly bounded

variances. Given $c \geq 1$, let ℓ denote the integer part of $c \log p$, and put

$$V_j' = \sum_{k=1}^{\ell} \alpha_{jk}\, \xi_{j-k}\,.$$

Then, by Markov's inequality,

$$P(|V_j - V_j'| > u) \leq u^{-2}\, E\Big(\sum_{k=\ell+1}^{\infty} \alpha_{jk}\, \xi_{j-k} \Big)^2 = O\big(u^{-2}\, \rho^{2\ell}\big)\,,$$

uniformly in $u > 0$, $c \geq 1$ and integers j. By taking $u = p^{-C}$ for $C > 0$ arbitrarily large, and then choosing $c \geq \frac{3}{2}\, C\, |\log \rho|^{-1}$, we deduce that the approximants V_j' have the following two properties: (a) $P(|V_j - V_j'| \leq p^{-C}) = 1 - O(p^{-C})$, uniformly in $1 \leq j \leq p$; and (b) for each r in the range $2 \leq r \leq p$, and each sequence $1 \leq j_1 < \ldots < j_r \leq p$ satisfying $j_{k+1} - j_k \geq c \log p + 1$ for $1 \leq k \leq r-1$, the variables V_{j_k}', for $1 \leq k \leq r$, are stochastically independent.

The regularity condition (B1), below, captures this behaviour in greater generality. There, we let V_{Wj}, for $1 \leq j \leq p$, have the joint distribution of the respective values of $Z_j - \bar{W}_{\cdot j}$ when Z is drawn from Π_W. At (B2), we also impose (a) a uniform Hölder smoothness condition on the respective distribution functions χ_{Wj} of V_{Wj}, (b) a symmetry condition on χ_{Wj}, and (c) a restriction which prevents the upper tail of Ψ_{Wj}, for each j and W, from being pathologically heavy.

Condition B

(B1) For each $C, \epsilon > 0$, and each of the two choices of W, there exists a sequence of random variables V_{Wj}', for $1 \leq j \leq p$, with the properties: (a) $P(|V_{Wj} - V_{Wj}'| \leq p^{-C}) = 1 - O(p^{-C})$, uniformly in $1 \leq j \leq p$; and (b) for all sufficiently large p, for each r in the range $2 \leq r \leq p$, and each sequence $1 \leq j_1 < \ldots < j_r \leq p$ satisfying $j_{k+1} - j_k \geq p^{\epsilon}$ for $1 \leq k \leq r-1$, the variables V_{Wj_k}', for $1 \leq k \leq r$, are stochastically independent.

(B2) (a) For each of the two choices of W there exist constants $C_1, C_2 > 0$, the former small and the latter large, such that $|\chi_{Wj}(u_1) - \chi_{Wj}(u_2)| \leq C_2\, |u_1 - u_2|^{C_1}$, uniformly in $u_1, u_2 > 0$, $1 \leq j \leq p < \infty$ and $W = X$ or Y; (b) the function $G_{Wj}(u, v) = P(|V_{Wj} + v| \leq u)$ is nonincreasing in $|v|$ for each fixed u, each choice of W and each j; and (c) $\max_{1 \leq j \leq p} \{1 - \Psi_{Wj}(u)\} = O(u^{-\epsilon})$, for $W = X, Y$ and for some $\epsilon > 0$.

Part (b) of (B2) holds if each distribution of V_{Wj} is symmetric and unimodal.

As explained in the previous section, in both the hypothesis testing and classification problems we can consider that $W = X$ or Y, indicating the population from which we draw the sample against which we check the new data value Z. Let $\mu_Z = \mu_X$ if Z is from Π_X, and $\mu_Z = \mu_Y$ otherwise, and define $\nu_{WZj} = \mu_{Zj} - \mu_{Wj}$,

$$\overline{\mathrm{hc}}_{WZ}(C,t) = \sup_{u \in \mathcal{U}_W(C,t)} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ P(|V_{Wj}| \le u) - P(|V_{Wj} + \nu_{WZj}| \le u) \right\}. \tag{3.3}$$

In view of (B2)(b), the quantity within braces in the definition of $\overline{\mathrm{hc}}_{WZ}$ is nonnegative, and so $\overline{\mathrm{hc}}_{WZ} \ge 0$. Theorem 3.2 below describes the extent to which the statistic hc_W, a random variable, can be approximated by the deterministic quantity $\overline{\mathrm{hc}}_{WZ}$.

Theorem 3.2. *Let $C > 0$ be fixed, and take the threshold, $t = t(p)$, to satisfy $t \ge 0$ and $\psi_W(t) \ge C$, thus ensuring that $\mathcal{U}_W(C,t)$ is nonempty. Let hc_W and $\overline{\mathrm{hc}}_{WZ}$ denote $\mathrm{hc}_W(C,t)$ and $\overline{\mathrm{hc}}_{WZ}(C,t)$, respectively. If* (B1) *and* (B2) *hold then for each $\epsilon > 0$,*

$$\mathrm{hc}_W = -\{1 + o_p(1)\}\,\overline{\mathrm{hc}}_{WZ} + O_p\big(p^{\epsilon}\big)\,. \tag{3.4}$$

An attractive feature of (3.4) is that it separates the "stochastic" and "deterministic" effects of the higher-criticism statistic hc_W. The stochastic effects go into the term $O_p(p^\epsilon)$. The deterministic effects are represented by $\overline{\mathrm{hc}}_{WZ}$. When the data value Z is from the same population Π_W as the dataset with which it is compared, each $\nu_{WZj} = 0$ and so, by (3.3), $\overline{\mathrm{hc}}_{WZ} = 0$. Property (3.4) therefore implies that, when Z is from Π_W, $\mathrm{hc}_W = O_p(p^\epsilon)$ for each $\epsilon > 0$. In other cases, where Z is drawn from a population different from that from which come the data with which Z is compared, $\overline{\mathrm{hc}}_{WZ}$ is generally nonzero. In such instances the properties of hc_W can be computed by relatively straightforward, deterministic calculations based on $\overline{\mathrm{hc}}_{WZ}$. In particular, when $W \ne Z$, if $\overline{\mathrm{hc}}_{WZ}$ is of order larger than p^ϵ for some $\epsilon > 0$ (see (3.8) below), then it follows directly that the probability of rejecting the null hypothesis, in the hypothesis testing problem, or of correct classification, in the classification problem, converges to 1. See, for example, section 3.3.

Remark 3.1: *Sharpening the term $O_p(p^\epsilon)$ in* (3.4). If, as in the problem treated by Donoho and Jin (2004), the distribution functions Ψ_{Wj} are all identical and the variables $X_{i_1 j_1}$ and $Y_{i_2 j_2}$, for $1 \le i_1 \le n_X$, $1 \le i_2 \le n_Y$ and $1 \le j_1, j_2 \le p$, are completely independent, then a refinement of the argument leading to (3.4) shows that the $O_p(p^\epsilon)$ term there can be reduced to $O_p(\sqrt{\log p})$. Here it is not necessary to assume that the common distributions are normal. Indeed, in that context Donoho and Jin (2004) noted that the $O_p(p^\epsilon)$ term in (3.4) can be replaced by $O_p(\sqrt{\log \log p})$.

Remark 3.2: *Relaxing the monotonicity condition* (B2)(b). Assumption (B2)(b) asks that $G_{Wj}(u, v) = P(|V_{Wj} + v| \le u)$ be nonincreasing in $|v|$ for each u. However, if the distributions of X_{ij} and Y_{ij} are identical for all but at most q values of j then it is sufficient to ask that, for these particular j, it be possible to write, for each $\epsilon > 0$,

$$G_{Wj}(u, v) = H_{Wj}(u, v) + o\{p^\epsilon \, q^{-1} \, \psi_W(u)^{1/2}\}\,,$$

uniformly in $1 \le j \le p$, $u \ge t$ and $W = X$ and Y, where each H_{Wj} has the monotonicity property asked of G_{Wj} in (B2)(b).

Remark 3.3: *Mixing at polynomial rate.* The exponential-like mixing rate implied by (B1) is a consequence of the fact that (a) and (b) in (B1) hold for each $C, \epsilon > 0$. If, instead, those properties apply only for a particular positive pair C, ϵ, then (3.4) continues to hold with p^ϵ there replaced by p^η, where $\eta > 0$ depends on C, ϵ from (B1), and decreases to zero as C increases and ϵ decreases.

3.3 Delineating good performance

Theorem 3.2 gives a simple representation of the higher-criticism statistic. It implies that, if Z is drawn from Π_Q where $(W, Q) = (X, Y)$ or (Y, X), and if $\overline{\text{hc}}_{WZ}$ exceeds a constant multiple of p^ϵ for some $\epsilon > 0$, then the probability that we make the correct decision in either a hypothesis testing or classification problem, Z converges to 1 as $p \to \infty$. We shall use this result to determine a region where hypothesis testing, or classification, are possible. For simplicity, in this section we shall assume that each $\mu_{Xj} = 0$, and $\mu_{Yj} = 0$ for all but q values of j, for which $\mu_{Yj} = \nu > 0$ and $\nu = \nu(p)$ diverges with p and does not depend on j. The explicit form of

$\overline{\text{hc}}_{WZ}$, at (3.3), makes it possible to handle many other settings, but a zero-or-ν representation of each mean difference permits an insightful comparison with results discussed by Donoho and Jin (2004).

In principle, two cases are of interest, where the tails of the distribution of V_{Wj} decrease polynomially or exponentially fast, respectively. However, in the polynomial case it can be proved using (3.3) that the hypothesis testing and classification problems are relatively simple. Therefore, we study only the exponential setting. In this context, and reflecting the discussion in section 2.2, we take the distribution of V_{Wj} to be that of the difference between two Studentised means, standardised by dividing by $\sqrt{2}$, and the distributions of X_{ij} and Y_{ij} to represent translations of that distribution. See (C1) and (C2) below. Alternatively we could work with the case where X_{ij} is a Studentised mean for a distribution with zero mean, and Y_{ij} is computed similarly but for the case where the expected value is shifted by $\pm\nu$. Theoretical arguments in the latter setting are almost identical to those given here, being based on results of Wang and Hall (2007).

Condition C

(C1) For each pair (W, j), where $W = X$ or Y and $1 \le j \le p$, let U_{Wjk}, for $1 \le k \le N_{Wj}$, denote random variables that are independent and identically distributed as U_{Wj}, where $E(U_{Wj}) = 0$, $E(U_{Wj}^4)$ is bounded uniformly in (W, j), $E(U_{Wj}^2)$ is bounded away from zero uniformly in (W, j), and $N_{Wj} \ge 2$. Let V_{Wj} have the distribution $2^{-1/2}$ times the difference between two independent copies of $N_{Wj}^{1/2}\, \bar{U}_{Wj}/S_{Wj}$, where $\bar{U}_{Wj}$ and S_{Wj} denote respectively the empirical mean and variance of the data $U_{Wj1}, \ldots, U_{WjN_{Wj}}$. Take $\mu_{Xj} = 0$ for each j, $\mu_{Yj} = 0$ for all but $q = q(p)$ values of j, say $j_1, \ldots, j_q$, and $|\mu_{Yj}| = \nu$ for these particular values of j.

(C2) The quantity ν in (C1) is given by $\nu = \sqrt{2w\, \log p}$, and the threshold, t, satisfies $B \le t \le \sqrt{2s\, \log p}$ for some $B, s > 0$, where $0 < w < 1$ and $0 < s < \min(4w, 1)$.

The setting described by (C1) is one where a working statistician would, in practice, generally take each distribution approximation $\widehat{\Psi}_{Wj}(u)$ to be simply $P(|\xi| \le u)$, where ξ has the standard normal distribution. The signal detection boundary in this setting is obtained using a polynomial model for the number of added shifts:

$$\text{for some} \quad \tfrac{1}{2} < \beta < 1\,, \quad q \sim \text{const.}\, p^{1-\beta} \tag{3.5}$$

(Donoho and Jin, 2004). The boundary is then determined by:

$$w \geq \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{if } \frac{3}{4} < \beta < 1\,. \end{cases} \tag{3.6}$$

The inequality (3.6) is also sufficient in hypothesis testing and classification problems where the data are exactly normally distributed. Likewise it is valid if we use a normal approximation and if that approximation is good enough. The question we shall address is, "how good is good enough?" The following theorem answers this question in cases where N_{Wj} diverges at at least a logarithmic rate, as a function of p. The proof is given is given in Appendix A.2.

Theorem 3.3. *Assume* (C1), (C2), (3.5), *that w satisfies* (3.6), *and that, for $W = X$ or Y and $1 \leq j \leq p$, N_{Wj}, given in* (C1), *satisfies*

$$N_{Wj}^{-1} \, (\log p)^4 \to 0\,. \tag{3.7}$$

Suppose too that Z is from Π_Q, where $(W, Q) = (X, Y)$ or (Y, X). Then, for constants $B, \eta > 0$,

$$\overline{\mathrm{hc}}_{WZ} \geq B\, p^{\eta}\,. \tag{3.8}$$

Condition (3.7) confirms that the samples on which the coordinate data are based need be only logarithmically large, as a function of p, in order for the higher-criticism classifier to be able to detect the difference between the W and Q populations.

4 Further results

4.1 Alternative constructions of hc_W and $\widehat{\mathrm{hc}}_W$

There are several other ways of constructing higher-criticism statistics when the distribution functions Ψ_{Wj} depend on j and have to be estimated. For example, omitting for simplicity the threshold t, we could re-define $\widehat{\mathrm{hc}}_W$ as:

$$\widehat{\mathrm{hc}}_W = p^{1/2} \inf_{u\,:\,pu(1-u) \geq C} \{u\,(1-u)\}^{-1/2} \sum_{j=1}^{p} \left[I\left\{ |Z_j - \bar{W}_{\cdot j}| \leq \widehat{\Psi}_{Wj}^{-1}(u) \right\} - u \right]. \tag{4.1}$$

If Z were drawn from Π_W then the random variable $K = \sum_j I\{|Z_j - \bar{W}_{\cdot j}| \le \Psi^{-1}_{Wj}(u)\}$ would have exactly a binomial $\mathrm{Bi}(p,u)$ distribution. The normalisation in formula (4.1) for $\widehat{\mathrm{hc}}_W$ reflects this property. However, replacing K by $\widehat{K} = \sum_j I\{|Z_j - \bar{W}_{\cdot j}| \le \widehat{\Psi}^{-1}_{Wj}(u)\}$, as in (4.1), destroys the independence of the summands, and makes normalisation problematic. This is particularly true when, as would commonly be the case in practice, the estimators $\widehat{\Psi}_{Wj}$ are computed from data W_{ij_1} for values of j_1 that are local to j. In such cases the estimators $\widehat{\Psi}_{Wj}$ would not be root-p consistent for the respective distributions Ψ_{Wj}.

If the distribution of $|Z_j - \bar{W}_{\cdot j}|$ were known up to its standard deviation, σ_{Wj}; and if we had an estimator, $\hat{\sigma}_{Wj}$, of σ_{Wj} for each W and j; then we could construct a third version of $\widehat{\mathrm{hc}}_W$:

$$\widehat{\mathrm{hc}}_W = \inf_{u\,:\,\phi_W(u) \ge C} \phi_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ I\big(|Z_j - \bar{W}_{\cdot j}|/\hat{\sigma}_{Wj} \le u\big) - \Phi_{Wj}(u) \right\},$$

where Φ_{Wj} denotes the distribution function of $|Z_j - \bar{W}_{\cdot j}|/\sigma_{Wj}$ under the assumption that Z is drawn from Π_W, and $\phi_W(u) = \sum_j \Phi_{Wj}\,(1 - \Phi_{Wj})$. Again, however, the correlation induced through estimation, this time the estimation of σ_{Wj}, makes the normalisation difficult to justify.

In some problems there is good reason to believe that if the marginal means of the populations Π_X and Π_Y differ, then the differences are of a particular sign. For example, it might be known that $\mu_{Xj} \ge \mu_{Yj}$ for all j. In this case we would alter the construction of the higher-criticism statistics hc_W and $\widehat{\mathrm{hc}}_W$, at (2.2) and (2.4), to:

$$\mathrm{hc}^{\mathrm{os}}_W = \inf_{u \in \mathcal{U}^{\mathrm{os}}_W} \psi^{\mathrm{os}}_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ I(Z_j - \bar{W}_{\cdot j} \le u) - \Psi^{\mathrm{os}}_{Wj}(u) \right\}, \tag{4.2}$$

$$\widehat{\mathrm{hc}}^{\mathrm{os}}_W = \inf_{u \in \widehat{\mathcal{U}}^{\mathrm{os}}_W} \widehat{\psi}^{\mathrm{os}}_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ I(Z_j - \bar{W}_{\cdot j} \le u) - \widehat{\Psi}^{\mathrm{os}}_{Wj}(u) \right\}, \tag{4.3}$$

respectively, where

$$\psi^{\mathrm{os}}_W(u) = \sum_{j=1}^{p} \Psi^{\mathrm{os}}_{Wj}(u) \left\{1 - \Psi^{\mathrm{os}}_{Wj}(u)\right\}, \quad \widehat{\psi}^{\mathrm{os}}_W(u) = \sum_{j=1}^{p} \widehat{\Psi}^{\mathrm{os}}_{Wj}(u) \left\{1 - \widehat{\Psi}^{\mathrm{os}}_{Wj}(u)\right\},$$

$\widehat{\Psi}^{\rm os}_{Wj}(u)$ is an empirical approximation to the probability $\Psi^{\rm os}_{Wj}(u) = P(Z_j - \bar{W}_{\cdot j} \le u)$ when Z is drawn from Π_W, $\mathcal{U}^{\rm os}_W = \mathcal{U}^{\rm os}_W(C,t)$ is the set of u for which $u \ge t$, $\psi^{\rm os}_W(u) \ge C$, $\widehat{\mathcal{U}}^{\rm os}_W$ is defined analogously, and the superscript "os" denotes "one-sided." When using $\widehat{\rm hc}^{\rm os}_W$ we would classify Z as coming from Π_X if $\widehat{\rm hc}^{\rm os}_X \ge \widehat{\rm hc}^{\rm os}_Y$, and as coming from Π_Y otherwise.

Remark 4.1: *Adapting Theorems 3.1 and 3.2.* Theorems 3.1 and 3.2 have direct analogues, formulated in the obvious manner, for the one-sided classifiers ${\rm hc}^{\rm os}_W$ and ${\rm hc}^{\rm os}_W$ introduced above. In particular, the one-sided version of $\overline{\rm hc}_{WZ}$, at (3.3), is obtained by removing the absolute value signs there. The regularity conditions too differ in only minor respects. For example, when formulating the appropriate version of (A1) we replace $\widehat{\Psi}_{Wj}$, Ψ_{Wj}, $\widehat{\psi}_W$ and ψ_W by $\widehat{\Psi}^{\rm os}_{Wj}$, $\Psi^{\rm os}_{Wj}$, $\widehat{\psi}^{\rm os}_W$ and $\psi^{\rm os}_W$, respectively. Part (b) of (B2) can be dropped on this occasion, since its analogue in the one-sided case follows directly from the monotonicity of a distribution function.

4.2 Advantages of incorporating the threshold

By taking the threshold, t, large we can construct the higher-criticism statistics ${\rm hc}_W$ and $\widehat{\rm hc}_W$, at (2.2) and (2.4), so that they emphasise relatively large values of $|Z_j - \bar{W}_{\cdot j}|$. This is potentially advantageous, especially when working with $\widehat{\rm hc}_W$, since we expect the value of u at which the infimum at (2.4) is achieved also to be large.

The most important reasons for thresholding are more subtle than this argument would suggest, however. They are founded on properties of relative errors in distribution approximations, and on the fact that the divisor in (2.2) is $\psi_W^{1/2}$, not simply ψ_W. To see why this is significant, consider the case where the distribution functions Ψ_{Wj} are all identical, to Ψ say. Then $\psi_W = p\,\Psi\,(1-\Psi)$, which we estimate by $\widehat{\psi}_W = p\,\widehat{\Psi}\,(1-\widehat{\Psi})$, say. In order for the effect of replacing each $\Psi_{Wj}(u)$ (appearing in (2.2)) by $\widehat{\Psi}_{Wj}(u)$ (in (2.4)) to be asymptotically negligible, we require the quantity

$$\psi_W(u)^{-1/2} \sum_{j=1}^{p} \bigl|\widehat{\Psi}_{Wj}(u) - \Psi_{Wj}(u)\bigr| = \frac{p^{1/2}|\widehat{\Psi}(u) - \Psi(u)|}{\Psi(u)^{1/2}\{1-\Psi(u)\}^{1/2}}$$

to be small. Equivalently, if u is in the upper tail of the distribution Ψ, we need

the ratio

$$\frac{p^{1/2}\,|\widehat{\Psi}(u) - \Psi(u)|}{\{1 - \Psi(u)\}^{1/2}} \tag{4.4}$$

to be small.

If the approximation of Ψ by $\widehat{\Psi}$ (or more particularly, of $1 - \Psi$ by $1 - \widehat{\Psi}$) is accurate in a relative sense, as it is (for example) if Ψ is the distribution of a Studentised mean, then, for large u,

$$\rho(u) \equiv \frac{|\widehat{\Psi}(u) - \Psi(u)|}{1 - \Psi(u)} \tag{4.5}$$

is small for u in the upper tail as well as for u in the middle of the distribution. When u is in the upper tail, so that $1 - \Psi(u)$ is small, then, comparing (4.4) and (4.5), we see that we do not require $\rho(u)$ to be as small as it would have to be in the middle of the distribution. By insisting that $u \geq t$, where the threshold t is relatively large, we force u to be in the upper tail, thus obtaining the advantage mentioned in the previous sentence.

Below, we show in more detail why, if thresholding is not undertaken, that is, if we do not choose t large when applying the higher-criticism classifier, substantial errors can occur when using the classifier. They arise through an accumulation of errors in the approximation $\widehat{\Psi}_{Wj} \approx \Psi_{Wj}$.

Commonly, the approximation of Ψ_{Wj} by $\widehat{\Psi}_{Wj}$ can be expressed as

$$\widehat{\Psi}_{Wj}(u) = \Psi_{Wj}(u) + \delta_p\,\alpha_{Wj}(u) + o(\delta_p)\,, \tag{4.6}$$

where δ_p decreases to zero as p increases and represents the accuracy of the approximation; α_{Wj} is a function, which may not depend on j; and the remainder, $o(\delta_p)$, denotes higher-order terms. Even if α_{Wj} depends on j, its contribution cannot be expected to "average out" of $\widehat{\mathrm{hc}}_W$, by some sort of law-of-large-numbers effect, as we sum over j.

In some problems the size of δ_p is determined by the number of data used to construct $\widehat{\Psi}_{Wj}$. For example, in the analysis of gene-expression data, $\widehat{\Psi}_{Wj}$ might be calculated by borrowing information from neighbouring values of j. In order for this method to be adaptive, only a small proportion of genes would be defined as

neighbours for any particular j, and so a theoretical description of δ_p would take that quantity to be no smaller than $p^{-\eta}$, for a small constant $\eta > 0$. In particular, assuming that $\widehat{\Psi}_{Wj}$ was root-p consistent for Ψ_{Wj}, i.e. taking η as large as $\frac{1}{2}$, would be out of the question.

In other problems the coordinate data X_{ij} and Y_{ij} can plausibly be taken as approximately normally distributed, since they are based on Student's t statistics. See sections 2.2 and 3.3 for discussion. In such cases the size of δ_p is determined by the number of data in samples from which the t statistic is computed. This would also be much less than p, and so again a mathematical account of the size of δ_p would have it no smaller than $p^{-\eta}$, for $\eta > 0$ much less than $\frac{1}{2}$.

Against this background; and taking, for simplicity, $\Psi = \Psi_{Wj}$, $\alpha = \alpha_W$ and $\delta_p = p^{-\eta}$; we find that $\widehat{\psi}_W \sim \psi_W = p\,\Psi\,(1-\Psi)$ and $\sum_j (\widehat{\Psi}_{Wj} - \Psi_{Wj}) = p^{1-\eta}\,\alpha + o(p^{1-\eta})$. These results, and (4.6), lead to the conclusion that, for fixed u, the argument of the infimum in the definition of $\widehat{\mathrm{hc}}_W$, at (2.4), is given by

$$\begin{aligned} A(u) &\equiv \widehat{\psi}_W(u)^{-1/2} \sum_{j=1}^{p} \big\{ I(|Z_j - \bar{W}_{\cdot j}| \le u) - \widehat{\Psi}_{Wj}(u) \big\} \\ &= \{1 + o(1)\}\, \psi_W(u)^{-1/2} \sum_{j=1}^{p} \big\{ I(|Z_j - \bar{W}_{\cdot j}| \le u) - \Psi_{Wj}(u) \big\} \\ &\qquad - p^{(1/2)-\eta}\, \gamma(u) + o\big(p^{(1/2)-\eta}\big)\,, \end{aligned} \tag{4.7}$$

where $\gamma = \alpha\,\{\Psi\,(1-\Psi)\}^{-1/2}$.

Assume, again for simplicity, that Z is drawn from Π_W. Then, for fixed u, the series on the right-hand side of (4.7) has zero mean, and equals $O_p(p^{1/2})$. In consequence,

$$A(u) = O_p(1) - p^{(1/2)-\eta}\, \gamma(u) + o_p\big(p^{(1/2)-\eta}\big)\,. \tag{4.8}$$

Referring to the definition of $\widehat{\mathrm{hc}}_W$ at (2.4), we conclude from (4.8) that for fixed u,

$$\widehat{\mathrm{hc}}_W \le O_p(1) - p^{(1/2)-\eta}\, \gamma(u) + o_p\big(p^{(1/2)-\eta}\big)\,. \tag{4.9}$$

If u is chosen so that $\gamma(u) > 0$ then, since $\eta < \frac{1}{2}$, the subtracted term on the right-hand side of (4.9) diverges to $-\infty$ at a polynomial rate, and this behaviour is

readily mistaken for detection of a value of Z that does not come from Π_W. (There, the rate of divergence to zero can be particularly small; see section 3.3 and Donoho and Jin (2004).) This difficulty has arisen through the accumulation of errors in the distribution approximation.

5 Numerical properties in the case of classification

We applied the higher-criticism classifier to simulated data. In each case, we generated $n_W = 10$ vectors of dimension $p = 10^6$, from $\Pi_W = \Pi_X$ or Π_Y; and one observation Z from Π_Y. We generated the data such that, for $i = 1, \ldots, n_W$ and $j = 1, \ldots, p$; and with W denoting X or Y; $W_{i,j} = (\bar{U}^W_{ij,1} - \bar{U}^W_{ij,2})/\sqrt{(S^2_{U,1} + S^2_{U,2})/N_U} + \mu_W$ where, for $k = 1$ and 2, $\bar{U}^W_{ij,k}$ was the empirical mean and $S^2_{U,k}$ was the empirical variance of: (1) in the case of independence, $N_U = 20$ independent and identically distributed random variables, having the distribution function of a $\mathrm{N}(0, 1)$ variable, a student T_{10} or a χ^2_6 random variable; and (2) in the case of dependence, $N_U = 20$ random variables of the type $V^W_{i,j,k}$, where, for $i = 1, \ldots, n_W$ and $j = 1, \ldots, p$, $V^W_{i,j,k} = \sum_{\ell=0}^{L} \theta^\ell \, \varepsilon^W_{i,j-\ell,k}$, with $\theta = 0.8$ and $\varepsilon^W_{i,j,k} \sim \mathrm{N}(0, (1 + \theta^2)^{-1})$ denoting independent variables.

We set $\mu_{X,j} = 5(j-1)/(p-1)$ and, in compliance with (C2), (3.5), (3.6), took $\mu_X = \mu_Y$ for all but $q = \langle p^{1-\beta} \rangle$ randomly selected components, for which $\mu_{Y_j} = \mu_{X,j} + \sqrt{2w \log p}$, where $\langle \cdot \rangle$ denotes the integer-part function; and we considered different values of $\beta \in (\frac{1}{2}, 1)$ and $w \in (0, \frac{1}{2})$. Reflecting the results in sections 3.1 and 3.3, we estimated the unknown distribution function of the observed data as the standard normal distribution function. In all cases considered, we generated 500 samples in the manner described above, and we repeated the classification procedure 500 times. Below we discuss the percentages of those samples which led to correct classification.

Application of the method necessitated selection of the two parameters t and C defining $\mathcal{U}_W$. In view of condition (A2), we reformulated $\mathcal{U}_W$ as $\mathcal{U}_W = [t_1, t_2]$, and we replaced choice of t and C by choice of t_1 and t_2. If we have sufficient experience with the distributions of the data, t_1 and t_2 can be selected 'theoretically' to maximise the percentage of correct classifications.

In the tables below we compare the results obtained using three methods: the higher-criticism procedure for the optimal choice of (t_1, t_2), referring to it as HC_T; higher criticism without thresholding, i.e. for $(t_1, t_2) = (-\infty, \infty)$, to which we refer as simply HC; and the thresholded nearest-neighbour method, NN_T (see e.g. Hall *et al.* 2008), i.e. the nearest-neighbour method applied to thresholded data $W_j\, I\{W_j > t\}$, where W denotes X, Y or Z and the threshold, t, is selected in a theoretically optimal way using the approach described above for choosing (t_1, t_2).

It is known (Hall *et al.*, 2008) that, for normal variables, when the distribution of the observations is known, classification using HC_T is possible if w and β are above the boundary determined by (3.6), but classification using NN_T is possible only above the more restricted boundary determined by $w = 2\beta - 1$. Below, we show that these results hold in our context too, where the distribution of the data is known only approximately (more precisely, estimated by the distribution of a standard normal variable). We shall consider values of (β, w) that lie above, on or below the boundary $w = 2\beta - 1$, including values which lie between this boundary and that for higher criticism. Tables 1 and 2 summarise results for the independent case (1), when the observations were averages of, respectively, Student T_{10} variables and χ^2_6 variables. In all cases, including those where classification was possible for both methods, we see that the thresholded higher-criticism method performs significantly better than the thresholded nearest-neighbour approach. The results also show very clearly the improvement obtainable using the thresholded version of higher criticism.

Table 1: *Percentage of correct classifications if case (1) with T_{10} variables, using the optimal values of t, t_1 and t_2.*

	$w = 0.2$			$w = 0.3$			$w = 0.4$			$w = 0.5$		
β	NN_T	HC_T	HC	NN_T	HC_T	HC	NN_T	HC_T	HC	NN_T	HC_T	HC
0.5	99.8	100	95.8									
0.6	77.4	86.4	83.0	86.2	97.4	89.6	94.0	100	94.0			
0.65	63.8	72.8	70.6	74.4	85.4	73.4	77.2	97.8	82.6			
0.7				63.2	70.6	62.2	67.6	86.0	64.6	69.8	96.8	75.8
0.75							59.0	72.8	58.2	65.2	86.2	66.6

Table 2: *Percentage of correct classifications if case (1) with χ^2_6 variables, using the optimal values of t, t_1 and t_2.*

	$w = 0.2$			$w = 0.3$			$w = 0.4$			$w = 0.5$		
β	NN_T	HC_T	HC	NN_T	HC_T	HC	NN_T	HC_T	HC	NN_T	HC_T	HC
0.5	99.8	100	97.6									
0.6	75.8	85.4	78.6	86.2	98.4	89.0	93.6	100	91.8			
0.65	66.2	70.2	68.0	69.8	86.2	72.4	80.6	98.2	79.2			
0.7				64.4	74.8	64.2	68.4	88.4	64.4	75.0	97.0	69.8
0.75							60.0	73.8	56.0	64.2	85.4	59.4

Table 3: *Percentage of correct classifications if case (1) with normal variables (line 1), case (2) with $L = 1$ (line 2) or $L = 3$ (line 3), using the optimal values of t, t_1 and t_2.*

	$w = 0.2$			$w = 0.3$			$w = 0.4$			$w = 0.5$		
β	NN_T	HC_T	HC	NN_T	HC_T	HC	NN_T	HC_T	HC	NN_T	HC_T	HC
0.5	99.8	100	94.0									
	98.0	100	93.8									
	95.6	100	93.6									
0.6	77.2	83.4	79.2	85.2	98.0	88.6	93.0	100	95.2			
	73.0	83.2	80.0	81.8	97.2	85.6	90.0	100	92.4			
	67.4	82.0	77.2	75.6	97.6	84.0	85.6	100	92.0			
0.65	63.6	70.2	65.6	71.0	83.8	74.2	79.4	97.3	82.4			
	61.8	72.0	68.4	68.2	84.0	72.4	76.0	97.8	82.8			
	58.2	67.8	64.4	62.6	83.0	72.0	73.2	96.4	79.2			
0.7				64.2	70.2	65.0	69.0	83.0	66.4	72.8	95.2	75.6
				59.6	69.6	57.4	63.6	84.2	67.2	71.2	95.0	76.6
				59.4	70.8	63.6	63.0	82.8	63.8	66.4	94.4	70.2
0.75							59.4	69.2	60.8	62.0	85.4	63.2
							59.0	71.4	59.2	59.8	85.0	62.6
							54.0	71.4	57.4	60.4	80.8	62.6

In Table 3 we compare the results of the independent case (1), where the data were Studentised means of independent $N(0, 1)$ variables and so had Student's t distribution; and the dependent case (2), where the observations were Studentised means of correlated normal variables with either $L = 1$ or $L = 3$. Here we see that as the strength of correlation increases, the nearest-neighbour method seems

to deteriorate more rapidly than higher criticism, which, as indicated in section 3.2, remains relatively unaffected by lack of independence.

If previous experience with data of the type being analysed is not sufficient to permit effective choice of threshold using that background, then a data-driven selection needs to be developed. We implemented a cross-validation procedure, described in Appendix A.1.

6 Technical arguments

6.1. Proof of Theorem 3.1. Since $\psi_W(u) \geq C$ for each $u \in \mathcal{U}_W(C,t)$ then (A1) implies that $\psi_W(u)^{-1} \sum_j |\widehat{\Psi}_{Wj}(u) - \Psi_{Wj}(u)| = o_p(1)$ uniformly in $u \in \mathcal{U}_W(C,t)$, and hence that $\widehat{\psi}_W(u)/\psi_W(u) = 1 + o_p(1)$, uniformly in $u \in \mathcal{U}_W(C,t)$. Call this result R_1. That property and (A2) imply that with probability converging to 1 as $p \to \infty$, $\mathcal{U}_W(C_3,t) \subseteq \widehat{\mathcal{U}}_W(C_2,t) \subseteq \mathcal{U}_W(C_1,t)$; call this result R_2. (Since $\psi_W(t) \geq C_3$ then $t \in \mathcal{U}_W(C_3,t)$, and so the latter set is nonempty.) Results R_1, R_2 and (A1) together give (3.1). Property R_2 and (3.1) imply (3.2).

6.2. Proof of Theorem 3.2. Let V'_{Wj} be as in (B1). Since, in the case where Z is drawn from Π_W, V_{Wj}, for $1 \leq j \leq p$, have the joint distribution of $Z_j - \bar{W}_{\cdot j}$, for $1 \leq j \leq p$, then for Z from either Π_X or Π_Y we may write $Z_j - \bar{W}_{\cdot j} = V_{Wj} + \nu_{WZj}$, where $\nu_{WZj} = \mu_{Zj} - \mu_{Wj}$. Substituting this representation for $Z_j - \bar{W}_{\cdot j}$ into the definition of hc_W at (2.2), and defining $\Delta_{Wj} = V_{Wj} - V'_{Wj}$, we see that

$$\mathrm{hc}_W = \inf_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \{ I(|V'_{Wj} + \Delta_{Wj} + \nu_{WZj}| \leq u) - \Psi_{Wj}(u) \} \,. \tag{6.1}$$

Given $D > 0$ and $v = 0$ or ± 1, define

$$\mathrm{hc}'_{WZ}(v) = \inf_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ I\big(|V'_{Wj} + \nu_{WZj}| \leq u + v\, p^{-D}\big) - \Psi_{Wj}(u) \right\},$$

$$\mathrm{hc}''_{WZ} = \inf_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \left\{ I(|V'_{Wj} + \nu_{WZj}| \leq u) - P(|V'_{Wj}| \leq u) \right\}.$$

Let $\mathcal{E}_W$ denote the event that $|\Delta_{Wj}| \leq p^{-D}$ for each $1 \leq j \leq p$. In view of (B1)(a),

$$\text{for all} \quad C_3 > 0 \,, \quad P(\mathcal{E}_W) = 1 - O\big(p^{-C_3}\big) \,. \tag{6.2}$$

Now, with probability 1,

$$\begin{aligned} &\mathrm{hc}'_{WZ}(-1) \le \mathrm{hc}'_{WZ}(0) \le \mathrm{hc}'_{WZ}(1) \qquad \text{and} \\ &\mathrm{hc}'_{WZ}(-1) \le \mathrm{hc}_W \le \mathrm{hc}'_{WZ}(1) \text{ if } \mathcal{E}_W \text{ holds}\,, \end{aligned} \tag{6.3}$$

where we used (6.1) to obtain the second set of inequalities. Furthermore,

$$\begin{aligned} 0 &\le \mathrm{hc}'_{WZ}(1) - \mathrm{hc}'_{WQ}(-1) \\ &\le \sup_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} I\big(u - p^{-D} < |V'_{Wj} + \nu_{WZj}| \le u + p^{-D}\big) \\ &\le \sup_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} I\big(u - 2p^{-D} < |V_{Wj} + \nu_{WZj}| \le u + 2p^{-D}\big)\,, \end{aligned} \tag{6.4}$$

where the first inequality holds with probability 1 and the second holds almost surely on $\mathcal{E}_W$.

Let $1 \le j_1 \le p$, and take $C_4 > 0$. Using (B1) and (B2) it can be shown that the probability that there are no integers $j_2 \ne j_1$ with $1 \le j_2 \le p$ and

$$\Big| |V_{Wj_1} + \nu_{Zj_1}| - |V_{Wj_2} + \nu_{Zj_2}| \Big| \le C_4\, p^{-D}\,, \tag{6.5}$$

is bounded below by $1 - C_5\, p^{1-DC_1}$ uniformly in j_1, where $C_5 > 0$ and C_1 is the constant in (B2)(a). Adding over $1 \le j_1 \le p$, and choosing $D > 2C_1^{-1}$, we deduce that:

> The probability that there is no pair (j_1, j_2) of distinct indices such that $|V_{Wj_1} + \nu_{Qj_1}|$ and $|V_{Wj_2} + \nu_{Qj_2}|$ are closer than $C_4\, p^{-D}$, converges to zero as $p \to \infty$. (6.6)

If, in the case $C_4 = 4$, the inequality (6.5) fails for all distinct integer pairs (j_1, j_2) with $1 \le j_1, j_2 \le p$, then the series on the far right-hand side of (6.4) can have no more than one nonzero term. That term, if it exists, must equal 1. In this case the far right-hand side of (6.4) cannot exceed $\sup_{u \in \mathcal{U}_W} \psi_W(u)^{-1/2}$, which in turn is bounded above by a constant, $C_6 = C^{-1/2}$. Hence, (6.4) and (6.6) imply that

$$P\big\{0 \le \mathrm{hc}'_{WZ}(1) - \mathrm{hc}'_{WZ}(-1) \le C_6\big\} \to 1\,. \tag{6.7}$$

Combining (6.2), (6.3) and (6.7) we deduce that

$$P\big\{|\mathrm{hc}_W - \mathrm{hc}'_{WZ}(0)| \le C_6\big\} \to 1\,. \tag{6.8}$$

Observe too that, uniformly in u,

$$\begin{aligned}\big|P(|V'_{Wj}| \le u) - \Psi_{Wj}(u)\big| &\le \big|\Psi_{Wj}\big(u + p^{-D}\big) - \Psi_{Wj}\big(u - p^{-D}\big)\big| + P(\mathcal{E}_W) \\ &\le C_7\left(4p^{-D}\right)^{C_1} + P(\mathcal{E}_W) = O\big(p^{-DC_1}\big)\,,\end{aligned}$$

where we have used (B2) and (6.2). Hence,

$$|\mathrm{hc}''_{WZ} - \mathrm{hc}'_{WZ}(0)| \le \sup_{u\in\mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \big|P(|V'_{Wj}| \le u) - \Psi_{Wj}(u)\big| = O\big(p^{-DC_1}\big)\,.$$

Combining this result and (6.8) we deduce that if $C_8 > 0$ is chosen sufficiently large,

$$P\big(|\mathrm{hc}_W - \mathrm{hc}''_{WZ}| \le C_8\big) \to 1\,. \tag{6.9}$$

Next we introduce further notation, defining $\Psi_{WZj}(u) = P(|V_{Wj} + \nu_{WZj}| \le u)$,

$$\begin{aligned}\Psi^{\mathrm{dash}}_{Wj}(u) &= P(|V'_{Wj}| \le u)\,, \qquad & \Psi^{\mathrm{dash}}_{WZj}(u) &= P(|V'_{Wj} + \nu_{WZj}| \le u)\,, \\ \psi_{WZ} &= \sum_{j=1}^{p} \Psi_{WZj}\,(1 - \Psi_{WZj})\,, & \psi^{\mathrm{dash}}_{WZ} &= \sum_{j=1}^{p} \Psi^{\mathrm{dash}}_{WZj}\,\big(1 - \Psi^{\mathrm{dash}}_{WZj}\big)\,, \\ \phi_{WZ} &= \sum_{j=1}^{p} (\Psi_{Wj} - \Psi_{WZj})\,, & \omega_{WZ} &= \psi_W + \phi_{WZ}\,, \\ \mathrm{hc}^{(3)}_{WZ} &= \sup_{u\in\mathcal{U}_W} \omega_{WZ}(u)^{-1/2} \left|\sum_{j=1}^{p} \big\{I(|V'_{Wj} + \nu_{WZj}| \le u) - \Psi^{\mathrm{dash}}_{WZj}(u)\big\}\right|, \\ \mathrm{hc}^{(4)}_{WZ} &= \sup_{u\in\mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \big\{P(|V'_{Wj}| \le u) - P(|V'_{Wj} + \nu_{WZj}| \le u)\big\} \\ &= \sup_{u\in\mathcal{U}_W} \psi_W(u)^{-1/2} \sum_{j=1}^{p} \big\{\Psi^{\mathrm{dash}}_{Wj}(u) - \Psi^{\mathrm{dash}}_{WZj}(u)\big\}\,.\end{aligned}$$

The remainder of the proof develops approximations to $\mathrm{hc}^{(3)}_{WZ}$ and $\mathrm{hc}^{(4)}_{WZ}$.

Using (B1)(a) and (B2)(a) it can be shown that, uniformly in u,

$$\big|\psi_{WZ} - \psi^{\mathrm{dash}}_{WZ}\big| = O(p^{1-DC_1}) \to 0\,, \tag{6.10}$$

provided $D > C_1^{-1}$. Also, if $D > C_1^{-1}$ then a similar argument can be used to show that, with $\overline{\mathrm{hc}}_{WZ}$ defined as at (3.3),

$$\bigl|\mathrm{hc}_{WZ}^{(4)} - \overline{\mathrm{hc}}_{WZ}\bigr| \to 0\,. \tag{6.11}$$

By (B2)(b), $0 \le \Psi_{WZj} \le \Psi_{Wj} \le 1$, from which it follows that

$$\begin{aligned}\Psi_{Wj}\,(1-\Psi_{Wj}) + \Psi_{Wj} - \Psi_{WZj}& \\ &= \Psi_{WZj}\,(1-\Psi_{WZj}) + (\Psi_{Wj} - \Psi_{WZj})\,(2 - \Psi_{Wj} - \Psi_{WZj}) \\ &\ge \Psi_{WZj}\,(1-\Psi_{WZj})\end{aligned}$$

for each j. Adding over j we deduce that $\omega_{WZ} \ge \psi_{WZ}$. Combining this result with (6.10), and noting that $\omega_{WZ}(u) \ge \psi_W(u) > C$ for $u \in \mathcal{U}_W = \mathcal{U}_W(C,t)$, we deduce that, for a constant $C_9 > 0$,

$$\text{for all } u \in \mathcal{U}_W\,, \quad \psi_{WZ}^{\mathrm{dash}}(u) \le C_9\,\omega_{WZ}(u)\,. \tag{6.12}$$

Write $\langle\cdot\rangle$ for the integer-part function. Given $\epsilon \in (0,1)$, use (B1)(b) to break the sum inside the absolute value in the definition of $\mathrm{hc}_{WZ}^{(3)}$, taken over $1 \le j \le p$, into $\langle p^\epsilon\rangle$ series, each consisting only of independent terms. Let $S_{WZk}(u)$, for $1 \le k \le \langle p^\epsilon\rangle$, denote the kth of these series. Now, $E\{S_{WZk}(u)\} = 0$ and, for $u \in \mathcal{U}_W$,

$$\mathrm{var}\{S_{WZk}(u)\} \le \psi_{WZ}^{\mathrm{dash}}(u) \le C_9\,\omega_{WZ}(u)\,, \tag{6.13}$$

where the variance is computed using the expression for $S_{WZk}(u)$ as a sum of independent random variables, and the second inequality comes from (6.12).

Employing (6.13), and noting again the independence property, standard arguments can be used to show that for each choice of $C_{10}, C_{11} > 0$,

$$\max_{1\le k\le\langle p^\epsilon\rangle} P\biggl\{\sup_{u\in\mathcal{U}_W} \omega_{WZ}(u)^{-1/2}\,|S_{WZk}(u)| > p^{C_{10}}\biggr\} = O\bigl(p^{-C_{11}}\bigr)\,. \tag{6.14}$$

In particular, using Rosenthal's inequality, Markov's inequality and the fact that $\omega_{WZ}(u) \ge \psi_W(u) \ge C$ for $u \in \mathcal{U}_W$, we may show that for all $B_1, B_2 > 0$,

$$\max_{1\le k\le\langle p^\epsilon\rangle} \sup_{u\in\mathcal{U}_W} P\bigl\{|S_{WZk}(u)| > \omega_{WZ}(u)^{1/2}\,p^{B_1}\bigr\} = O\bigl(p^{-B_2}\bigr)\,.$$

Therefore, if $\mathcal{V}_W = \mathcal{V}_W(p)$ denotes any subset of $\mathcal{U}_W$ that contains only $O(p^{B_3})$ elements, for some $B_3 > 0$, then for all $B_1, B_2 > 0$,

$$\max_{1\le k\le\langle p^\epsilon\rangle} P\Big\{ \max_{u\in\mathcal{V}_W} \omega_{WZ}(u)^{-1/2}\,|S_{WZk}(u)| > p^{B_1} \Big\} = O\big(p^{B_3-B_2}\big) = O\big(p^{-B_4}\big)\,, \tag{6.15}$$

where $B_4 = B_2 - B_3$. Since B_3 and B_4 both can be taken arbitrarily large, then, using the monotonicity of the function $g(u) = I(v \le u)$, and also properties (B1) and (B2), it can be seen that $\max_{u\in\mathcal{V}_W}$ in (6.15) can be replaced by $\sup_{u\in\mathcal{U}_W}$, giving (6.14). In this context, condition (B2)(c) ensures that, with an error that is less than p^{-B_5}, for any given $B_5 > 0$, the distribution Ψ_{Wj} can be truncated at a point p^{B_6}, for sufficiently large B_6; and, within the interval $[0, p^{B_6}]$, the points in $\mathcal{V}_W$ can be chosen less than p^{-B_7} apart, where $B_7 > 0$ is arbitrarily large.

Result (6.14) implies that

$$P\Big\{ \max_{1\le k\le\langle p^\epsilon\rangle} \sup_{u\in\mathcal{U}_W} \omega_{WZ}(u)^{-1/2}\,|S_{WZk}(u)| > p^{C_{10}} \Big\} = O\big(p^{\epsilon-C_{11}}\big)\,,$$

from which it follows that $P(\mathrm{hc}^{(3)}_{WZ} > p^{\epsilon+C_{10}}) = O(p^{\epsilon-C_{11}})$. Since ϵ, C_{10} and C_{11} are arbitrary positive numbers then we may replace ϵ here by zero, obtaining: for each $C_{10}, C_{11} > 0$,

$$P\big(\mathrm{hc}^{(3)}_{WZ} > p^{C_{10}}\big) = O\big(p^{-C_{11}}\big)\,. \tag{6.16}$$

It can be deduced directly from the definitions of hc''_{WZ}, $\mathrm{hc}^{(3)}_{WZ}$ and $\mathrm{hc}^{(4)}_{WZ}$ that:

$$|\mathrm{hc}''_{WZ} + \mathrm{hc}^{(4)}_{WZ}| \le \mathrm{hc}^{(3)}_{WZ} \sup_{u\in\mathcal{U}_W} \left\{ \frac{\omega_{WZ}(u)}{\psi_W(u)} \right\}^{1/2} = \mathrm{hc}^{(3)}_{WZ} \sup_{u\in\mathcal{U}_W} \left\{ 1 + \frac{\phi_{WZ}(u)}{\psi_W(u)} \right\}^{1/2}.$$

Combining this result with (6.9), (6.11) and (6.16); and noting that

$$\overline{\mathrm{hc}}_{WZ} = \sup_{u\in\mathcal{U}_W} \frac{\phi_{WZ}(u)}{\psi_W(u)^{1/2}}\,,$$

and, since $\psi_W(u) \ge C$ for $u \in \mathcal{U}_W$,

$$\sup_{u\in\mathcal{U}_W} \left\{ 1 + \frac{\phi_{WZ}(u)}{\psi_W(u)} \right\}^{1/2} \le \big(1 + C^{-1/4}\big) \sup_{u\in\mathcal{U}_W} \left\{ 1 + \frac{\phi_{WZ}(u)}{\psi_W(u)^{1/2}} \right\}^{1/2};$$

we deduce that for each $\epsilon > 0$,

$$\mathrm{hc}_W + \overline{\mathrm{hc}}_{WZ} = O_p\Big\{p^{\epsilon}\left(1 + \overline{\mathrm{hc}}_{WZ}\right)^{1/2}\Big\}. \tag{6.17}$$

Theorem 3.2 follows directly from (6.17).

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D.L. AND JOHNSTONE, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653.

BENJAMANI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* Ser. B **57**, 289–300.

BROBERG, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biol.* **4**, R41 (electronic).

CAI, T., JIN, J. AND LOW, M. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, to appear.

CAYÓN, L., BANDAY, A.J., JAFFE, T., ERIKSEN, H.K.K., HANSEN, F.K., GORSKI, K.M. AND JIN, J. (2006). No higher criticism of the Bianchi corrected WMAP data. *Mon. Not. Roy. Astron. Soc.* **369**, 598–602.

CAYÓN, L., JIN, J. AND TREASTER, A. (2005). Higher criticism statistic: Detecting and identifying non-Gaussianity in the WMAP first year data. *Mon. Not. Roy. Astron. Soc.* **362**, 826–832.

CRUZ, M., CAYÓN, L., MARTÍNEZ-GONZÁLEZ, E., VIELVA, P. AND JIN, J. (2007). The non-Gaussian cold spot in the 3-year WMAP data. *Astrophys. J.* **655**, 11–20.

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. Amer. Statist. Assoc.* **99**, 96–104.

FAN, J., CHEN, Y., CHAN, H.M., TAM, P., AND REN, Y. (2005a). Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells. *Proc. Nat. Acad. Sci. USA* **102**, 17751–17756.

FAN, J. AND FAN, Y. (2007). High dimensional classification using features annealed independence rules. *Ann. Statist.*, to appear.

FAN, J., PENG, H., AND HUANG, T. (2005b). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. (With discussion.) *J. Amer. Statist. Assoc.* **100**, 781–813.

FAN, J., TAM, P., VANDE WOUDE, G. AND REN, Y. (2004). Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Nat. Acad. Sci. USA* **101**, 1135–1140.

HALL, P. AND JIN, J. (2006). Properties of higher criticism under long-range dependence. *Ann. Statist.*, to appear.

HALL, P., PITTELKOW, Y. AND GHOSH, M. (2008). On relative theoretical performance of classifiers suitable for high-dimensional data and small sample sizes. *J. Roy. Statist. Soc.* Ser. B, to appear.

HUANG, J., WANG, D. AND ZHANG, C. (2003). A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. Manuscript.

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

INGSTER, Yu. I. (2001). Adaptive detection of a signal of growing dimension. I. Meeting on Mathematical Statistics. *Math. Methods Statist. 10*, 395–421.

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

INGSTER, Yu. I. AND SUSLINA, I.A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer, New York.

JIN, J. (2006). Higher criticism statistic: Theory and applications in non-Gaussian detection. In: *Statistical Problems in Particle Physics, Astrophysics And Cosmology*, Eds. L. Lyons and M.K. Ünel. Imperial College Press, London.

JIN, J., STARCK, J.-L., DONOHO, D.L., AGHANIM, N. AND FORNI, O. (2004). Cosmological non-Gaussian signature detection: Comparing performance of different statistical tests. *Eurasip J. Appl. Signal Processing* **15**, 2470–2485.

LEHMANN, E.L., ROMANO, J.P. AND SHAFFER, J.P. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33**, 1084–1108.

LÖNNSTEDT, I. AND SPEED, T. (2002). Replicated microarray data. *Statist.*

Sinica **12**, 31–46.

MEINSHAUSEN, M. AND RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independent tested hypotheses. *Ann. Statist.* **34**, 373–393.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

STOREY, J.D. AND TIBSHIRANI, R. (2003). Statistical significance for genome-wide experiments. *Proc. Nat. Acad. Sci. USA* **100**, 9440–9445.

STOREY, J.D., TAYLOR, J.E., AND SIEGMUND, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Stat. Soc.* Ser. B **66**, 187–205.

TSENG, G.C., OH, M.K., ROHLIN, L., LIAO, J.C. AND WONG, W.H. (2001). Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549–2557.

TUKEY, J.W. (1953). The problem of multiple comparisons. Manuscript. Department of Statistics, Princeton University.

WANG, Q. (2005). Limit theorems for self-nomalized large deviation. *Electronic J. Probab.* **38**, 1260–1285.

WANG, Q AND HALL, P. (2007). Relative errors in central limit theorem for Student's t statistic, with applications. Manuscript.

APPENDIX

A.1. Description of the cross-validation procedure. If previous experience with data of the type being analysed is not sufficient to permit effective choice of threshold using that background, then a data-driven selection needs to be developed. This, however, is a challenging task, as the sample sizes are typically very small. As a first practical method, we implemented a cross-validation (CV) procedure where the basic idea was as follows. Create $n_X + n_Y$ cross-validation samples $(\mathcal{X}_{CV,k}, \mathcal{Y}_{CV,k}, Z_{CV,k}) = (\mathcal{W}^{(-j)}, \mathcal{T}, W_j)$, $k = j + n_X\, I(W = Y)$, $j = 1, \ldots, n_W$, $(W, T) = (X, Y)$ or (Y, X), where $\mathcal{W}^{(-j)}$ denotes the sample $\mathcal{W}$ with the jth observation W_j left out; apply the classification procedure to each CV sample, and then choose (t_1, t_2) to give a large number of correct classifications, but not too large so

as to avoid 'overfitting' the data. We experimented with different ways of avoiding the overfitting problem, and found that the following gave quite good results.

(a) Here we describe how to choose the grid on which we search for (t_1, t_2). One of the problems in our context is that p is so large that removing one of the data values, as is done in cross-validation, has substantial impact on the range of the observed data. Therefore, and since we expect t_2 to be related to the extreme observed values, it would not be appropriate to choose a grid for (t_1, t_2) and keep it fixed over each iteration of the algorithm. Instead, at each step k, where $k = 1, \ldots, n_X + n_Y$, of the algorithm we define the grid in terms of a set of $K \in [2, 2p-1]$ order statistics $U_{(i_1)} < U_{(i_2)} < \ldots < U_{(i_K)}$ of the vector $U = (|Z_{CV,k} - \bar{X}_{CV,k}|, |Z_{CV,k} - \bar{Y}_{CV,k}|)$. (To make notations less heavy, we omit the index k from U.) We keep fixed the vector $I = (i_1, \ldots, i_K)$ of K indices. At each step we define our grid for (t_1, t_2) as $U_{(I)} \times U_{(I)}$, where $U_{(I)}$ denotes $(U_{(i_1)}, \ldots, U_{(i_K)})$. The indices $1 \le i_1 < i_2 < \ldots < i_K \le 2p$ are chosen such that the last, say, $K - S$ order statistics $V_{(i_{S+1})} < V_{(i_{S+2})} < \ldots < V_{(i_K)}$ of the vector $V = (|Z_k - \bar{X}_k|, |Z_k - \bar{Y}_k|)$ consist of the extreme values of V, and the first S order statistics $V_{(i_1)} < V_{(i_2)} < \ldots < V_{(i_S)}$ are uniformly distributed over the interval $[V_{(1)}, V_{(i_{S+1}-1)}]$.

(b) For $k = 1, \ldots, n_X + n_Y$, apply the HC procedure to the kth cross-validation sample, for each (t_1, t_2) in the grid $U_{(I)} \times U_{(I)}$.

(c) For each $1 \le j, k \le K$, let $C_{j,k}$ denote the number of correct classifications out of the $n_X + n_Y$ cross-validation trials at (b), obtained by taking $(t_1, t_2) = (U_{(i_j)}, U_{(i_k)})$. Of course, since t_1 must be less than t_2, we set $C_{j,k} = 0$ for all $j > k$.

(d) Taking V as in (a), construct the vector t_2^* of all values $V_{(i_k)}$ for which $\sup_j C_{j,k} \ge M' = \sup_{j,k} C_{j,k} - (n_X + n_Y)/10$. The factor $(n_X + n_Y)/10$ was chosen heuristically and it is introduced to avoid overfitting the data. Take t_2 as the component of t_2^*, say $V_{(i_\ell)}$, for which $\#\{j \text{ s.t. } C_{j,\ell} \ge M'\}$ is the largest — in case of non uniqueness, take $V_{(i_\ell)}$ as the largest such component. Then take t_1 as the average of all $V_{(i_j)}$'s such that $C_{j,\ell} \ge M'$.

In most cases this method gave good results, with performance lying approximately midway between that using the theoretically optimal (t_1, t_2) or no thresh-

olding, i.e. $(t_1, t_2) = (-\infty, \infty)$, respectively.

A.2. Proof of Theorem 3.3. For simplicity, we denote N_{Wj} by N. Recall that χ_{Wj} denotes the distribution of V_{Wj}, i.e. the distribution of $Z_j - \bar{W}_{\cdot j}$ when Z is drawn from Π_W. It can be proved from results of Wang (2005) that, under (C1), uniformly in values of $u > 0$ that satisfy $u = o(N^{-1/6})$, and uniformly in W and in $1 \le j \le p$,

$$\chi_{Wj}(u) = \Phi(u) + O\big[N^{-1/2}\,|u|^3\,\{1 - \Phi(u)\}\big]\,,$$

where Φ denotes the standard normal distribution function. An analogous result for $u \le 0$ also holds. Hence for $u > 0$ satisfying $u = o(N^{-1/6})$, we have uniformly in W and in $1 \le j \le p$,

$$P(|V_{Wj}| \le u) = 2\,\Phi(u) - 1 + O\big[N^{-1/2}\,u^3\,\{1 - \Phi(u)\}\big]\,. \tag{A.1}$$

Similarly it can be shown that, uniformly in $j = j_1, \ldots, j_q$, the latter as in (C1),

$$\begin{aligned} P(|V_{Wj} \pm \nu| \le u) = \Phi(u + \nu) + \Phi(u - \nu) - 1 \\ + O\big[N^{-1/2}\,(u + \nu)^3\,\{1 - \Phi(|u - \nu|)\}\big]\,. \end{aligned} \tag{A.2}$$

Let $a_{WQ}(u)$ denote the series in the definition of $\overline{\mathrm{hc}}_{WQ}$, at (3.3). Combining (A.1) and (A.2) we deduce that, if $Q \ne W$,

$$\begin{aligned} a_{WQ}(u) &= q\,\big\{2\,\Phi(u) - \Phi(u + \nu) - \Phi(u - \nu)\big\} \\ &\qquad + O\big[N^{-1/2}\,q\,(u + \nu)^3\,\{1 - \Phi(|u - \nu|)\}\big]\,, \qquad\qquad (A.3) \\ \psi_W(u) &= 2p\,\{2\Phi(u) - 1\}\,\{1 - \Phi(u)\} + O\big[N^{-1/2}\,p\,u^3\,\{1 - \Phi(u)\}\big] \\ &= \{1 + o(1)\}\,2p\,\{2\Phi(u) - 1\}\,\{1 - \Phi(u)\}\,, \qquad\qquad (A.4) \end{aligned}$$

uniformly in $u \in \mathcal{U}_W(C, t)$. To obtain the second identity in (A.4) we used the properties $t \ge B > 0$ and $N^{-1/2}\,(\log p)^{3/2} \to 0$, from (C2) and (3.7) respectively.

Take $u = \sqrt{2v \log p}$ where $0 < v = v(p) \le 1$, and recall that $\nu = \sqrt{2w\,\log p}$, where w and s are as in (C2). It can be shown, borrowing ideas from Donoho and Jin (2004), that

$$2\,\Phi(u) - \Phi(u + \nu) - \Phi(u - \nu) = g_1(p)\,p^{-(\sqrt{v} - \sqrt{w})^2}\,, \tag{A.5}$$

$$\{2\Phi(u) - 1\}\,\{1 - \Phi(u)\} \sim C_1\,(\log p)^{-1/2}\,p^{-v}\,, \tag{A.6}$$

where, here and below, g_j denotes a function that is bounded above by C_2 and below by $C_3\,(\log p)^{-1/2}$, and C_1, C_2, C_3 denote positive constants. To derive (A.5), write $2\,\Phi(u) - \Phi(u+\nu) - \Phi(u-\nu)$ as

$$\{1 - \Phi(u+\nu)\} + \{1 - \Phi(u-\nu)\} - 2\,\{1 - \Phi(u)\}\,,$$

and use conventional approximations to $1 - \Phi(z)$, for moderate to large positive z, and, when $u - \nu < 0$, to $\Phi(-z)$, for z in the same range.

In view of (A.5) and (A.6),

$$\frac{2\,\Phi(u) - \Phi(u+\nu) - \Phi(u-\nu)}{[2p\,\{2\Phi(u)-1\}\,\{1-\Phi(u)\}]^{1/2}} = g_2(p)\,(\log p)^{1/4}\,p^{b_1}\,, \tag{A.7}$$

where $b_1 = \frac{1}{2}\,(v-1) - (\sqrt{v} - \sqrt{w})^2$. Similarly,

$$\frac{N^{-1/2}\,(u+\nu)^3\,\{1 - \Phi(|u-\nu|)\}}{[p\,\{1-\Phi(u)\}]^{1/2}} = O\big\{N^{-1/2}\,(\log p)^{7/4}\,p^{b_1}\big\}\,. \tag{A.8}$$

Using (A.3), (A.4), (A.7) and (A.8) we deduce that, provided $N^{-1}\,(\log p)^4 \to 0$,

$$\frac{a_{WQ}(u)}{\psi_W(u)^{1/2}} \sim q\,g_2(p)\,(\log p)^{1/4}\,p^{b_1} = g_3(p)\,(\log p)^{1/4}\,p^{b_2}\,, \tag{A.9}$$

where $b_2 = \frac{1}{2}\,(v+1) - \beta - (\sqrt{v} - \sqrt{w})^2$.

Since s, in the definition of $t = \sqrt{2s\,\log p}$, satisfies $0 < s < \min(4w, 1)$, we can take

$$v = \begin{cases} 4w & \text{if } 0 < w < \frac{1}{4} \\ 1 - c(\log p)^{-1}\log\log p & \text{if } \frac{1}{4} \le w < 1\,, \end{cases}$$

where $c > \frac{1}{2}$, and have

$$u = \sqrt{2v\,\log p} = \min\big(2\nu, \sqrt{2\log p - 2c\log\log p}\big) \in \mathcal{U}(C,t)\,.$$

For this choice of v, $b_2 = 2\eta$ where $\eta > 0$, and it follows from (A.9) that

$$\overline{\mathrm{hc}}_{WQ} \ge \frac{a_{WQ}(u)}{\psi_W(u)^{1/2}} \ge C_4\,p^{\eta}\,.$$

Result (3.8) follows.