Robustness and accuracy of methods for high dimensional data analysis based on Student’s $t$-statistic

Aurore Delaigle,
*University of Melbourne, Australia*

Peter Hall
*University of Melbourne, Australia, and University of California at Davis, USA*

and Jiashun Jin
*Carnegie Mellon University, Pittsburgh, USA*

[Received January 2010. Revised September 2010]

Summary. Student’s $t$-statistic is finding applications today that were never envisaged when it was introduced more than a century ago. Many of these applications rely on properties, e.g. robustness against heavy-tailed sampling distributions, that were not explicitly considered until relatively recently. We explore these features of the $t$-statistic in the context of its application to very high dimensional problems, including feature selection and ranking, the simultaneous testing of many different hypotheses and sparse, high dimensional signal detection. Robustness properties of the $t$-ratio are highlighted, and it is established that those properties are preserved under applications of the bootstrap. In particular, bootstrap methods correct for skewness and therefore lead to second-order accuracy, even in the extreme tails. Indeed, it is shown that the bootstrap and also the more popular but less accurate $t$-distribution and normal approximations are more effective in the tails than towards the middle of the distribution. These properties motivate new methods, e.g. bootstrap-based techniques for signal detection, that confine attention to the significant tail of a statistic.

Keywords: Bootstrap; Central limit theorem; Classification; Dimension reduction; Higher criticism; Large deviation probability; Moderate deviation probability; Ranking; Second-order accuracy; Skewness; Tail probability; Variable selection

1. Introduction

Modern high throughput devices generate data in abundance. Gene microarrays comprise an iconic example; there, each subject is automatically measured on thousands or tens of thousands of standard features. What has not changed, however, is the difficulty of recruiting new subjects, with the number remaining in the tens or low hundreds. This is the context of so-called ‘$p \gg n$ problems’, where $p$ denotes the number of features, or the dimension, and $n$ is the number of subjects, or the sample size.

For each feature the measurements across different subjects comprise samples from potentially different underlying distributions and can have quite different scales and be highly skewed.
and heavy tailed. To standardize for scale, a conventional approach today is to use $t$-statistics, which, by virtue of the central limit theorem, are approximately normally distributed when $n$ is large. W. S. Gosset, when he introduced the Studentized $t$-statistic more than a century ago (Student, 1908), saw that quantity as having principally the virtue of scale invariance. In more recent times, however, other noteworthy advantages of Studentizing have been discovered. In particular, the $t$-statistic’s high degree of robustness against heavy-tailed data has been quantified. For example, Giné et al. (1997) have shown that a necessary and sufficient condition for the Studentized mean to have a limiting standard normal distribution is that the sampled distribution lies in the domain of attraction of the normal law. This condition does not require the sampled data to have finite variance. Moreover, the rate of convergence of the Studentized mean to normality is strictly faster than that for the conventional mean, normalized by its theoretical (rather than empirical) standard deviation, in cases where the second moment is only just finite (Hall and Wang, 2004). Contrary to the case of the conventional mean, its Studentized form admits accurate large deviation approximations in heavy-tailed cases where the sampling distribution has only a small number of finite moments (Shao, 1999).

All these properties are direct consequences of the advantages that are conferred by dividing the sample mean $\bar{X}$ by the sample standard deviation $S$. Erratic fluctuations in $\bar{X}$ tend to be cancelled, or at least dampened, by those of $S$, much more so than if $S$ were replaced by the true standard deviation of the population from which the data were drawn.

The robustness of the $t$-statistic is particularly useful in high dimensional data analysis, where the signal of interest is frequently found to be sparse. For any given problem (e.g. classification, prediction or multiple testing), only a small fraction of the automatically measured features are relevant. However, the locations of the useful features are unknown, and we must separate them empirically from an overwhelmingly large number of more useless features. Sparsity gives rise to a shift of interest away from problems involving vectors of conventional size to those involving high dimensional data.

As a result, a careful study of moderate and large deviations of the Studentized ratio is indispensable to understanding even common procedures for analysing high dimensional data, such as ranking methods based on $t$-statistics, or their applications to highly multiple-hypothesis testing (i.e. the simultaneous testing of many different hypotheses). See, for example, Benjamini and Hochberg (1995), Pigeot (2000), Finner and Roters (2002), Kesselman et al. (2002), Dudoit et al. (2003), Bernhard et al. (2004), Genovese and Wasserman (2004), Lehmann et al. (2005), Donoho and Jin (2006), Sarkar (2006), Jin and Cai (2007), Wu (2008), Kulinskaya (2009) and Cai and Jin (2010). The same issues arise in the case of methods for signal detection, e.g. those based on Student’s $t$-versions of higher criticism; see Donoho and Jin (2004), Jin (2008) and Delaigle and Hall (2009). Work in the context of multiple-hypothesis testing includes that of Lang and Secic (1997), page 63, Tamhane and Dunnett (1999), Takada et al. (2001), David et al. (2005), Fan et al. (2007) and Clarke and Hall (2009).

In the present paper we explore moderate and large deviations of the Studentized ratio in a variety of high dimensional settings. Our results reveal several advantages of Studentizing. We show that the bootstrap can be particularly effective in relieving skewness in the extreme tails. Attractive properties of the bootstrap for multiple-hypothesis testing were apparently first noted by Hall (1990), although in the case of the mean rather than its Studentized form.

Section 2.1 draws together several known results in the literature to demonstrate the robustness of the $t$-ratio in the context of high level exceedances. Sections 2.2 and 2.3 show that, even for extreme values of the $t$-ratio, the bootstrap captures particularly well the influence that departure from normality has on tail probabilities. We treat cases where the probability of
Robustness and Accuracy of Methods for High Dimensional Data Analysis

Section 2.4 shows how these properties can be applied to high dimensional problems, involving potential exceedances of high levels by many different feature components. One example of this type is the use of t-ratios to implement higher criticism methods, including their application to classification problems. This type of methodology is taken up in Section 3. The conclusions that are drawn in Sections 2 and 3 are illustrated numerically in Section 4, the underpinning theoretical arguments are summarized in Appendix A and detailed arguments are given by Delaigle et al. (2010).

2. Main conclusions and theoretical properties

2.1. Advantages and drawbacks of Studentizing in the normal approximation

Let $X_1, X_2, \ldots$ denote independent univariate random variables all distributed as $X$, with unit variance and zero mean, and suppose that we want to test $H_0: \mu = 0$ against $H_1: \mu > 0$. Two common test statistics for this problem are the standardized mean $Z_0$ and the Studentized mean $T_0$, which are defined by

$$Z_0 = \frac{n}{1 - 2} \bar{X},$$

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

denote the sample mean and sample variance respectively, computed from the data set $X_1, \ldots, X_n$.

In practice, experience with the context often suggests the standardization that defines $Z_0$. Although both $Z_0$ and $T_0$ are asymptotically normally distributed, dividing by the sample standard deviation introduces a degree of extra noise which can make itself felt in terms of greater effect of skewness. However, we shall show that, compared with the normal approximation to the distribution of $Z_0$, the normal approximation to the distribution of $T_0$ is valid under much less restrictive conditions on the tails of the distribution of $X$.

These properties will be established by exploring the relative accuracies of normal approximations to the probabilities $P(Z_0 > x)$ and $P(T_0 > x)$, as $x$ increases, and the conditions for validity of those approximations. This approach reflects important applications in problems such as multiple-hypothesis testing, and classification or ranking involving high dimensional data, since there it is necessary to assess the relevance, or statistical significance, of large values of sample means.

We start by showing that the normal approximation is substantially more robust for $T_0$ than it is for $Z_0$. To derive the results, note that if

$$E\{|X|^3\} < \infty$$

then the normal approximation to the probability $P(T_0 > x)$ is accurate, in relative terms, for $x$ almost as large as $n^{1/6}$. In particular, $P(T_0 > x)/\{1 - \Phi(x)\} \to 1$ as $n \to \infty$, uniformly in the range $0 < x \leq \epsilon n^{1/6}$, for any positive sequence $\epsilon$ that converges to 0 (Shao, 1999). This level of accuracy applies also to the normal approximation to the distribution of the non-Studentized mean $\bar{X}$, except that we must impose a condition that is much more severe than inequality (2.2). In particular, $P(Z_0 > x)/\{1 - \Phi(x)\} \to 1$, uniformly in $0 < x \leq \epsilon n^{1/6} - \eta$, for each fixed $\eta > 0$, if and only if

$$E\{\exp(|X|^c)\} < \infty$$

for all $c \in (0, \frac{1}{2})$;
see Linnik (1961). Condition (2.3), which requires exponentially light tails and implies that all moments of \( X \) are finite, is much more severe than condition (2.2).

Although dividing by the sample standard deviation confers robustness, it also introduces a degree of extra noise. To quantify deleterious effects of Studentizing we note that

\[
P(T_0 > x) = \{1 - \Phi(x)\}\{1 - \frac{1}{2}n^{-1/2}x^3 \gamma + o(n^{-1/2}x^3)\}, \tag{2.4}
\]

\[
P(Z_0 > x) = \{1 - \Phi(x)\}\{1 + \frac{1}{2}n^{-1/2}x^3 \gamma + o(n^{-1/2}x^3)\}, \tag{2.5}
\]

uniformly in \( x \) satisfying \( \lambda_n \leq x \leq n^{1/6} \lambda_n \), for a sequence \( \lambda_n \to \infty \), and where \( \Phi \) is the standard normal distribution function and \( \gamma = E(X^3) \) (Shao (1999) and Petrov (1975), chapter 8). (Property (2.2) is sufficient for equation (2.4) if \( x \to \infty \) and \( n^{-1/2}x^3 \to 0 \) as \( n \to \infty \), and equation (2.5) holds, for the same range of values of \( x \), provided that, for some \( u > 0 \), \( E\{\exp(u|X|)\} < \infty \).) Thus it can be seen that, if \( \gamma \neq 0 \) and \( n^{-1/2}x^3 \) is small, the relative error of the normal approximation to the distribution of \( T_0 \) is approximately twice that of the approximation to the distribution of \( Z_0 \).

Of course, Student’s \( t \)-distribution with \( n \) or \( n - 1 \) degrees of freedom is identical to the distribution of \( T_0 \) when \( X \) is normal \( N(0, \sigma^2) \) and therefore relates to the case of zero skewness. Taking \( \gamma = 0 \) in equation (2.4) we see that, when \( T_0 \) has Student’s \( t \)-distribution with \( n \) or \( n - 1 \) degrees of freedom, we have

\[
P(T_0 > x) = \{1 - \Phi(x)\}\{1 + o(n^{-1/2}x^3)\}.
\]

It can be deduced that the results derived in equations (2.4) and (2.5) continue to hold if we replace the role of the normal distribution by that of Student’s \( t \)-distribution with \( n \) or \( n - 1 \) degrees of freedom. Similarly, the results on robustness hold if we replace the role of the normal distribution by that of Student’s \( t \)-distribution. Thus, approximating the distributions of \( T_0 \) and \( Z_0 \) by that of a Student \( t \)-distribution, as is sometimes done in practice, instead of that of a normal distribution, does not alter our conclusions. In particular, even if we use the Student \( t \)-distribution, \( T_0 \) is still more robust against heavy tailedness than \( Z_0 \) and, in cases where the Student approximation is valid, this approximation is slightly more accurate for \( Z_0 \) than it is for \( T_0 \).

2.2. Correcting skewness by using the bootstrap

The arguments in Section 2.1 show clearly that \( T_0 \) is considerably more robust than \( Z_0 \) against heavy-tailed distributions, arguably making \( T_0 \) the test statistic of choice even if the population variance is known. However, as also shown in Section 2.1, this added robustness comes at the expense of a slight loss of accuracy in the approximation. For example, in equations (2.4) and (2.5) the main errors that arise in normal (or Student’s \( t \))- approximations to the distributions of \( T_0 \) are the result of uncorrected skewness. In the present section we show that if we instead approximate the distribution of \( T_0 \) by using the bootstrap then those errors can be quite successfully removed. Similar arguments can be employed to show that a bootstrap approximation to the distribution of \( Z_0 \) is less affected by skewness than is a normal approximation. However, as for the normal approximation, the bootstrap approximation is only valid if the distribution of \( X \) is very light tailed. Therefore, even if we use the bootstrap approximation, \( T_0 \) remains the statistic of choice.

Let \( X^* = \{X_1^*, \ldots, X_n^*\} \) denote a resample drawn by sampling randomly, with replacement, from \( X = \{X_1, \ldots, X_n\} \), and put
\[
\bar{X}^* = \frac{1}{n} \sum_{i=1}^{n} X_i^*, \\
S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)^2,
\]

\[
T_0^* = \frac{n^{1/2}(X_i^* - \bar{X})}{S^*}.
\]

The bootstrap approximation to the distribution function \( G(t) = P(T_0 \leq t) \) is \( \hat{G}(t) = P(T_0^* \leq t|X) \), and the bootstrap approximation to the quantile \( i_\alpha = (1 - G)^{-1}(\alpha) \) is

\[
i_\alpha = (1 - \hat{G})^{-1}(\alpha).
\]

Theorem 1, below, addresses the effectiveness of these approximations for large values of \( x \).

To appreciate why the bootstrap, when used appropriately, can be expected to correct for at least some of the effects of skewness, observe that if \( F \) denotes the distribution of a generic \( X_i \) then the distribution function \( G \) of \( T_0 \), which is defined below expression (2.6), can be written as \( G(t) = H(t|F) \), where \( H(\cdot|F) \) is a functional. In this notation the bootstrap distribution \( \hat{G} \) is just \( H(\cdot|\hat{F}) \), where \( \hat{F} \) is the empirical distribution function of the data. Now, the skewness of the distribution of \( X_i^* \) given the data, i.e. of the distribution \( \hat{F} \), is a consistent estimator of the skewness of the distribution of \( X_i \), and so skewness can be expected to be captured accurately by \( \hat{G} \) because it is inherited from \( \hat{F} \).

As usual in hypothesis testing problems, to calculate the level of the test we take a generic variable that has the distribution of the test statistic and we calculate the probability that the generic variable is larger than the estimated \( (1 - \alpha) \)-quantile. This generic variable is independent of the sample, and since the quantile \( \hat{i}_\alpha \) of the bootstrap test is random and constructed from the sample then, to avoid confusion, we should arguably use different notations for \( \hat{i}_\alpha \) and \( i_\alpha \), however to simplify the notation we keep using \( \hat{i}_\alpha \) denotes the random variable defined at equation (2.7) and calculated from the sample. In particular, here \( T_0 \) is independent of \( \hat{i}_\alpha \).

Define

\[
z_\alpha = (1 - \Phi)^{-1}(\alpha),
\]

and write \( P_F \) for the probability measure when \( X \) is drawn from the population with distribution function \( F \). Here we highlight the dependence of the probabilities on \( F \) because we shall use the results in subsequent sections where a clear distinction of the distribution will be required.

**Theorem 1.** For each \( B > 1 \) and \( D_1 > 0 \) there exists \( D_2 > 2 \), increasing no faster than linearly in \( D_1 \) as the latter increases, such that

\[
P_F(T_0 > \hat{i}_\alpha) = \alpha[1 + O\{(1 + z_\alpha)n^{-1/2} + (1 + z_\alpha)^4n^{-1}\}] + O(n^{-D_1})
\]

as \( n \to \infty \), uniformly in all distributions \( F \) of the random variable \( X \) such that \( E(|X|^{D_2}) \leq B(E(X)^2)^{D_2/2} \) and \( E(X) = 0 \), and in all \( \alpha \) satisfying \( 0 \leq z_\alpha \leq Bn^{1/4} \).

The theorem can be deduced by taking \( c = 0 \) in theorem 5 in Appendix A.1 and shows that using the bootstrap to approximate the distribution of \( T_0 \) removes the main effects of skewness. To appreciate why, note that if we were to use the normal approximation to the distribution of \( T_0 \) we would obtain, instead of equation (2.8), the following result, which can be deduced from theorem 4 in Appendix A.1 for each \( B > 1 \) such that \( E(|X|^4) < B \) and \( 0 \leq z_\alpha \leq Bn^{1/4} \):

\[
P_F(T_0 > z_\alpha) = \alpha \exp(-\frac{1}{3}n^{-1/2}z_\alpha^3\gamma)[1 + O\{(1 + z_\alpha)n^{-1/2} + (1 + z_\alpha)^4n^{-1}\}].
\]
Comparing equations (2.8) and (2.9) we see that the bootstrap approximation has removed the skewness term that describes first-order inaccuracies of the standard normal approximation.

The term in $(1 + z_\alpha)n^{-1/2}$ in equation (2.8) can be dropped if the distribution of $X$ is assumed to be sufficiently smooth. The version of equation (2.8) that results reflects second-order accuracy of the bootstrap; see, for example, Efron (1987) and Hall (1988). However, the term in $(1 + z_\alpha)n^{-1/2}$ cannot be dropped in the case of equation (2.9), since neither the normal approximation that is implicit in the use of $\exp(-\frac{1}{2}n^{-1/2}z_\alpha^2)$ on the right-hand side adequately compensates for skewness.

To appreciate that the $1 + O\{\cdots\}$ remainder term in equation (2.8) is of relatively minor importance in the problem of discovering large means in large $p$–small $n$ problems, note that the remainder can be removed by multiplying $\hat{t}_\alpha$ by a factor $1 + O(n^{-1/2}z_\alpha^{-1} + n^{-1}z_\alpha^2)$. However, since

$$\hat{t}_\alpha = z_\alpha[1 - \frac{1}{2}n^{-1/2}(1 + o_p(1))z_\alpha^2 \gamma],$$

where the skewness correction $-\frac{1}{2}n^{-1/2}(1 + o_p(1))z_\alpha^2 \gamma$ is of strictly larger order than $n^{-1/2}z_\alpha^{-1} + n^{-1}z_\alpha^2$, then an adjustment for the remainder in equation (2.8) would be minor relative to the skewness correction. In this sense the remainder term in equation (2.8), and in related formulae below, has a relatively minor effect on the process of knowledge discovery which, in most large $p$–small $n$ problems, motivates multiple-hypothesis testing. Importantly, the aim in such cases is not to test the intersection of all $p$ null hypotheses, but to rank a relatively small number of them in terms of the extent of evidence against them. These arguments show that the bootstrap that is applied to Student’s $t$-statistic can have a significant positive effect on our capacity for avoiding false discoveries by using methods such as those of Benjamini and Hochberg (1995), Blair et al. (1996), Storey (2002) and Genovese and Wasserman (2004).

The size of the $O(n^{-D_1})$ remainder in equation (2.8) is important if we wish to use the bootstrap approximation in the context of detecting $p$ weak signals, or of hypothesis testing for a given level of false discovery rate among $p$ populations or features. (Here and below it is convenient to take $p$ to be a function of $n$, which we treat as the main asymptotic parameter.) In all these cases we generally wish to take $\alpha$ of size $p^{-1}$, in the sense that $p\alpha$ is bounded away from 0 and $\infty$ as $n \to \infty$. This property entails $z_\alpha = O\{\log(p)^{1/2}\}$, and therefore theorem 1 implies that the tail condition $E(|X|^{D_2}) < \infty$, for some $D_2 > 0$, is sufficient for it to be true that

$$P_F(T_0 > \hat{t}_\alpha)/\alpha = 1 + o(1) \ \text{for} \ p = o(n^{D_1}) \ \text{and uniformly in the class of distributions}$$

$F$ of $X$ for which $E(X) = 0$ and $E(|X|^{D_2}) \leq B\{E(X^2)^2\}^{D_2/2}$.

In contrast, if, as in Fan and Lv (2008), $p$ is exponentially large as a function of $n$, then we require a finite exponential moment of $X$. The following theorem addresses this case. In the theorem, $D_2 < 2$ unless $D_1 = \frac{3}{8}$, in which case $D_2 = 2$. The proof of theorem 2 is given in Appendix A.2.

**Theorem 2.** For each $B > 1$ and $D_1 \in (0, \frac{3}{8})$ there exists $D_2 \in (0, 2]$, increasing no faster than linearly in $D_1$ as the latter increases, such that

$$P_F(T_0 > \hat{t}_\alpha) = \alpha[1 + O\{(1 + z_\alpha)n^{-1/2} + (1 + z_\alpha)^4n^{-1}\}] + O\{\exp(-n^{D_1})\} \quad (2.10)$$

as $n \to \infty$, uniformly in all distributions $F$ of the random variable $X$ such that $P\{|X| > x/E(X^2)^{1/2}\} \leq C\exp(-x^{D_2})$ (where $C > 0$) and $E(X) = 0$, and in all $\alpha$ satisfying $0 \leq z_\alpha \leq Bn^{1/4}$.

Theorem 2 allows us to repeat all the remarks that were made in connection with theorem 1 but in the case where $p$ is exponentially large as a function of $n$. Of course, we need to assume that exponential moments of $X$ are finite, but in return we can control a variety of statistical methodologies, such as sparse signal recovery or false discovery rate, for an exponentially large
number of potential signals or tests. Distributions with finite exponential moments include exponential families and distributions of variables that are supported on a compact domain. Our condition is still less restrictive than assuming that the distribution is normal, as is done in many references treating high dimensional problems, such as Fan and Lv (2008).

2.3. Effect of a non-zero mean on the properties discussed in Section 2.2

We have shown that, in a variety of problems, when making inference on a mean it is preferable to use the Studentized mean rather than the standardized mean. We have also shown that, when the skewness of the distribution of X is non-zero, the level of the test based on the Studentized mean is better approximated when using the bootstrap than when using a normal distribution. Our next task is to check that, when \( H_0 : \mu = 0 \) is not true, the probability of rejecting \( H_0 \) is not much affected by the bootstrap approximation. Our development is notionally simpler if we continue to assume that \( E(X) = 0 \) and \( \text{var}(X) = 1 \), and consider the test \( H_0 : \mu = -cn^{-1/2} \) with \( c > 0 \) a scalar that potentially depends on \( n \) but which does not converge to 0. We define

\[
Z_c = n^{1/2}(\bar{X} + cn^{-1/2}),
\]

\[
T_c = Z_c/S.
\]

(2.11)

Here we take \( \mu \) of magnitude \( n^{-1/2} \) because this represents the limiting case where inference is possible. Indeed, a population with mean of order \( o(n^{-1/2}) \) could not be distinguished from a population with mean 0. Thus we treat the statistically most challenging problem.

Our aim is to show that the probability \( P_F(T_c > t_\alpha) \) is well approximated by \( P_F(T_c > \hat{t}_\alpha) \), where \( c > 0 \) and \( \hat{t}_\alpha \) is given by equation (2.7), and when \( T_c \) and \( \hat{t}_\alpha \) are computed from independent data. We claim that in this setting the results that were discussed in Section 2.2 continue to hold.

In particular, versions of equations (2.8) and (2.10) in the present setting are

\[
P_F(T_c > \hat{t}_\alpha) = P_F(T_c > t_\alpha)[1 + O\{(1 + z_\alpha)n^{-1/2} + (1 + z_\alpha)^4n^{-1}\}] + R,
\]

(2.12)

where the remainder term \( R \) has either the form in equation (2.8) or that in equation (2.10), depending on whether we assume existence of polynomial or exponential moments respectively. In particular, if we take \( R = O(n^{-D_2}) \) then equation (2.12) holds uniformly in all distributions \( F \) of the random variable \( X \) such that \( E(|X|^{D_2}) \leq BE(X^2)^{D_2/2} \) and \( E(X) = 0 \), and in all \( \alpha \) satisfying \( 0 \leq z_\alpha \leq Bn^{1/4} \), provided that \( D_2 \) is sufficiently large; and in the same sense, but with \( R = O\{\exp(-n^{D_2})\} \) where \( D_1 \in (0, \frac{3}{8}] \), equation (2.12) holds if we replace the assumption \( E(|X|^{D_2}) \leq BE(X^2)^{D_2/2} \) by \( P\{|X| > x/E(X^2)^{1/2}\} \leq C\exp(-x^{D_2}) \), provided that \( D_2 \in (0, 2] \) is sufficiently large. (We require \( D_2 = 2 \) only if \( D_1 = \frac{3}{8} \).) Result (2.12) is derived in Appendix A.3. Hence, to first order, the probability of rejecting \( H_0 \) when \( H_0 \) is not true is not affected by the bootstrap approximation. In particular, to first order, skewness does not affect the approximation any more than it would if \( H_0 \) were true (compare with equation (2.8) and equation (2.10)).

An alternative form of equation (2.12), which is useful in applications (e.g. in Section 3), is to express the right-hand side there more explicitly in terms of \( \alpha \). This can be done if we note that, in view of theorem 4 in Appendix A.1,

\[
P_F(T_c > t_\alpha) = \{1 - \Phi(t_\alpha)\} \exp\{-\frac{1}{6}n^{-1/2}(2t_\alpha^3 - 3ct_\alpha^2 + c^3)\gamma\} \frac{1 - \Phi(t_\alpha - c)}{1 - \Phi(t_\alpha)}
\]

\[
\times [1 + \theta(c, n, t_\alpha)(1 + t_\alpha)n^{-1/2} + (1 + t_\alpha)^4n^{-1}]
\]

\[
= \alpha \exp\{\frac{1}{6}n^{-1/2}c(3t_\alpha^2 - c^2)\gamma\} \frac{1 - \Phi(t_\alpha - c)}{1 - \Phi(t_\alpha)}[1 + \theta_1(c, n, t_\alpha)(1 + t_\alpha)n^{-1/2}
\]

\[
+ (1 + t_\alpha)^4n^{-1}]
\]

(2.13)
where $\gamma$ denotes skewness, $\theta_j$ has the same interpretation as $\theta$ in theorem 4 and the last identity follows from the definition of $t_\alpha$. Combining this property with equation (2.12) it can be shown that

$$P_F(T_c > \hat{t}_\alpha) = \alpha \exp\left\{ \frac{1}{8} n^{-1/2} c(3t_\alpha^2 - c^2) \gamma \right\} \frac{1 - \Phi(t_\alpha - c)}{1 - \Phi(t_\alpha)} [1 + O\{(1 + z_\alpha)n^{-1/2} + (1 + z_\alpha)^4 n^{-1}\}] + R,$$

(2.14)

where $R$ satisfies the properties given below equation (2.12).

2.4. Relationships between many events $T_c > \hat{t}_\alpha$

So far we have treated only an individual event (i.e. a single univariate test), exploring its likelihood. However, since our results for a single event apply uniformly over many choices of the distribution of $X$ then we can develop properties in the context of many events, and thus for simultaneous tests. The simplest case is that where the values of $T_c$ are independent, i.e. we observe $T^{(j)}_c$ for $1 \leq j \leq p$, where $c^{(1)}, \ldots, c^{(p)}$ are constants and the random variables $T^{(j)}_c$ are, for different values $j$, computed from independent data sets. We assume that $T^{(j)}_c$ is defined as at expression (2.11) but with $c = c^{(j)}$. We could take the values of $n = n_j$ to depend on $j$, and in fact the theoretical discussion below remains valid provided that $C_1 n \leq n_j \leq C_2 n$, for positive constants $C_1$ and $C_2$, as $n$ increases. (Recall that $n$ is the main asymptotic parameter, and $p$ is interpreted as a function of $n$.) As in the case of a single event, which is treated in theorems 1 and 2, it is important that the $t$-statistic $T^{(j)}_c$ and the corresponding quantile estimator $\hat{t}^{(j)}_\alpha$ be independent for each $j$. However, as noted in Section 2.2, this is not a problem since $T^{(j)}_c$ represents a generic random variable, and only $t^{(j)}_\alpha$ is calculated from the sample.

It is often unnecessary to assume, as above, that the quantile estimators $\hat{t}^{(j)}_\alpha$ are independent of one another. To indicate why, we note that the method for deriving expansions such as equations (2.8), (2.10) and (2.12) involves computing $P(T_c > \hat{t}_\alpha)$ by first calculating the conditional probability $P(T_c > \hat{t}_\alpha | \hat{t}_\alpha)$, where the independence of $T_c$ and $\hat{t}_\alpha$ is used. Versions of this argument can be given for the case of short-range dependence between many different values of $\hat{t}^{(j)}_\alpha$, for $1 \leq j \leq p$.

Cases where the statistics are computed from weakly dependent data can be addressed by using results of Hall and Wang (2010). That work treats instances where the variables $T^{(j)}_c$ are computed from the first $n$ components in respective data streams $S_j = (X_{j1}, X_{j2}, \ldots)$, with $X_{j1}, X_{j2}, \ldots$ being independent and identically distributed but correlated between streams. As in the discussion above, since we are treating $t$-statistics then it can be assumed without loss of generality that the variables in each data stream have unit variance. (This condition serves only to standardize scale, and in particular places the means $c^{(j)}$ on the same scale for each $j$.) Assuming that this is so, we shall suppose also that third moments are uniformly bounded. Under these conditions it is shown by Hall and Wang (2010) that, provided that

(a) the correlations are bounded away from 1,
(b) the streams $S_1, S_2, \ldots$ are $k$ dependent for some fixed $k \geq 1$,
(c) $\gamma$ is bounded between two constant multiples of $\log(p)\sqrt{2}$,
(d) $\log(p) = o(n)$ and
(e) for $1 \leq j \leq p$ we have $0 \leq c^{(j)} = c^{(j)}(n) \leq \epsilon n^{-1/2} \log(p)^{1/2}$, where $\epsilon \to 0$ as $n \to \infty$,

and excepting realizations that arise with probability no greater than $1 - O\{p \exp(-Cz_\alpha^2)\}$, where $C > 0$, the $t$-statistics $T^{(j)}_c$ can be considered to be independent. In particular, it can be stated that with probability $1 - O\{p \exp(-Cz_\alpha^2)\}$ there are no clusters of level exceedances caused by dependence between the data streams.
These conditions, especially (d), permit the dimension $p$ to be exponentially large as a function of $n$. Assumption (e) is of interest; without it the result can fail and clustering can occur. To appreciate why, consider cases where the data streams are $k$ dependent but in the degenerate sense that $S_{rj} = \ldots = S_{rj+k}$ for $r \geq 0$. Then, for relatively large values of $c$, the value of $T_{c}^{(j)}$ is well approximated by that of $c/S_{j}$, where

$$S_{j}^{2} = n^{-1} \sum_{i<n} (X_{ij} - \bar{X}_{j})^{2}$$

is the empirical variance computed from the first $n$ data in the stream $S_{j}$. It follows that, for any $r \geq 1$, the values of $T_{c}^{(rj+i)}$, for $1 \leq i \leq k$, are also very close to one another. Clearly this can lead to data clustering that is not described accurately by asserting independence. There is evidence that, for genomic data, the strength of dependence can range from weak (Mansilla et al., 2004) to strong (Almirantis and Provata, 1999), and in the latter case the assumption of independence would be questionable.

![Fig. 1. Comparison of the joint distribution function of $T_{c}^{(1)}, \ldots, T_{c}^{(p)}$ (-----) with the product of the distributions of the univariate components $T_{0}^{(k)}$, $k = 1, \ldots, p$ (--------), when $\varepsilon_{k} \sim$ standardized Pareto(5,5), $n = 50$ and (a) $(p, \theta) = (100, 0.5)$, (b) $(p, \theta) = (100, 0.2)$ and (c) $(p, \theta) = (10000, 0.2)$: the vertical axis gives values of $P(T_{0}^{(1)} \leq x, \ldots, T_{0}^{(p)} \leq x)$ where $x$ is given on the horizontal axis](image-url)
To illustrate these properties we calculated the joint distribution of \((T_{01}^{(1)}, \ldots, T_{0}^{(p)})\) for short-range dependent \(p\)-vectors \((X_1, \ldots, X_p)\) and compared this distribution with the product of the distributions of the \(p\) univariate components \(T_{0k}^{(k)}, k = 1, \ldots, p\). For \(k = 1, \ldots, p\) we took \(X_k = \{U_k - E(U_k)\}/\sqrt{\text{var}(U_k)}\) and \(U_k = \sum_{j=0}^{10} \theta^j \epsilon_{j+k}\). Here, \(0 < \theta < 1\) is a constant and \(\epsilon_1, \ldots, \epsilon_{p+10}\) denote independent identically distributed random variables. Fig. 1 depicts the resulting distribution functions for several values of \(\theta\) and \(p\), when the sample size \(n\) was 50 and the \(\epsilon_j\)s were from a standardized Pareto(5,5) distribution. We see that the independence assumption gives a good approximation to the joint cumulative distribution function, but, unsurprisingly, the approximation degrades as \(\theta\) (and thus the dependence) increases. Fig. 1 also suggests that the independence approximation degrades as \(p\) becomes very large \((10^5\), in this example).

3. Application to higher criticism for detecting sparse signals in non-Gaussian noise

In this section we develop higher criticism methods where the critical points are based on bootstrap approximations to distributions of \(t\)-statistics, and we show that the advantages that were established in Section 2 for bootstrap \(t\)-methods carry over to sparse signal detection.

Assume that we observe \(X_{1j}, \ldots, X_{nj}\), for \(1 \leq j \leq p\), where all the observations are independent and where, for each \(j\), \(X_{1j}, \ldots, X_{nj}\) are identically distributed. For example, in gene microarray analysis \(X_{ij}\) is often used to represent the log-intensity that is associated with the \(i\)th subject and the \(j\)th gene, \(\mu_j\) represents the mean expression level associated with the \(j\)th feature (i.e. gene) and the \(Z_{ij}\)s represent measurement noise. The distributions of the \(X_{ij}\)s are completely unknown, and we allow the distributions to differ between components. Let \(E(X_{1j}) = c^{(j)}\). The problem of signal detection is to test

\[
H_0: \text{all } c^{(j)} \text{ are 0} \quad \text{against} \quad H_1^{(n)}: \text{a small fraction of the } c^{(j)} \text{ is non-zero.} \tag{3.1}
\]

For simplicity, in this section we assume that each \(c^{(j)} \geq 0\), but a similar treatment can be given where non-zero \(c^{(j)}\)s have different signs.

To perform the signal detection test we use the ideas in Section 2 to construct a bootstrap \(t\) higher criticism statistic that can be calculated when the distribution of the data is unknown, and which is robust against heavy tailedness of this distribution. (Higher criticism was originally suggested by Donoho and Jin (2004) in cases where the centred data have a known distribution, non-Studentized means were used and the bootstrap was not employed.) As in Section 2.4, let \(T_{c}^{(j)}\) be the Studentized statistic for the \(j\)th component, and let \(\hat{T}_{\alpha}^{(j)}\) be the bootstrap estimator of the \((1 - \alpha)\)-quantile of the distribution of \(T_{c}^{(j)}\), both calculated from the data \(X_{1j}, \ldots, X_{nj}\).

We suggest the following bootstrap \(t\) higher criticism statistic:

\[
\text{hc}_{n}(\alpha_0) = \max_{\alpha = i/p, 1 \leq i \leq \alpha_0 p} \left\{ p\alpha(1 - \alpha) \right\}^{-1/2} \sum_{j=1}^{p} \left\{ I(T_{c}^{(j)} > \hat{T}_{\alpha}^{(j)}) - \alpha \right\}, \tag{3.2}
\]

where \(\alpha_0 \in (0, 1)\) is sufficiently small for the statistic \(\text{hc}_{n}\) at equation (3.2) to depend only on indices \(j\) for which \(T_{c}^{(j)}\) is relatively large. This exploits the excellent performance of bootstrap approximation to the distribution of the Studentized mean in the tails, as exemplified by theorems 1 and 2 in Section 2, while avoiding the ‘body’ of the distribution, where the bootstrap approximations are sometimes less remarkable. We reject \(H_0\) if \(\text{hc}_{n}(\alpha_0)\) is too large.

We could have defined the higher criticism statistic by replacing the bootstrap quantiles in definition (3.2) by the respective quantiles of the standard normal distribution. However, the greater accuracy of bootstrap quantiles compared with normal quantiles, which was established in Section 2, suggests that in the higher criticism context, also, better performance can
be obtained when using bootstrap quantiles. The superiority of the bootstrap approach will be illustrated numerically in Section 4.

Theorem 3 below provides upper and lower bounds for the bootstrap $t$ higher criticism statistic at equation (3.2), under $H_0$ and $H_1^{(n)}$. We shall use these results to prove that the probabilities of type I and type II errors converge to 0 as $n \to \infty$. The standard ‘test pattern’ for assessing higher criticism is a sparse signal, with the same strength at each location where it is non-zero. It is standard to take $c^{(j)} = 0$ for all except a fraction $\varepsilon_n$ of $j$s, and $c^{(j)} = \tau_n n^{-1/2}$ elsewhere, where $\tau_n \neq 0$ is chosen to make the testing problem difficult but solvable. As usual in the higher criticism context we take

$$\varepsilon_n = p^{-\beta} = n^{-\beta/\theta},$$

where $\beta \in (0, 1)$ is a fixed parameter. Among these values of $\beta$ the range $0 < \beta < \frac{1}{2}$ is the least interesting, because there the proportion of non-zero signals is so high that it is possible to estimate the signal with reasonable accuracy, rather than just to determine its existence. See Donoho and Jin (2004). Therefore we focus on the most interesting range, which is $\frac{1}{2} < \beta < 1$.

For $\beta \in \left(\frac{1}{2}, 1\right)$ the most interesting values of $\tau_n$ are $\tau_n \approx \sqrt{\{2 \log(p)\}}$, with $\tau_n < \sqrt{\{2 \log(p)\}}$. Taking $\tau_n = [\alpha/\sqrt{\{2 \log(p)\}}]$ would render the two hypotheses indistinguishable, whereas taking $\tau_n \geq \sqrt{\{2 \log(p)\}}$ would render the signal relatively easy to discover, since it would imply that the means that are non-zero are of the same size as, or larger than, the largest values of the signal-free $T^{(j)}$s. In light of this we consider non-zero means of size

$$\tau_n = \sqrt{\{2r \log(p)\}} = \sqrt{\{2(r/\theta) \log(n)\}},$$

where $0 < r < 1$ is a fixed parameter.

Before stating the theorem we introduce notation. Let $L_p > 0$ be a generic multilog-term which may be different from one occurrence to the other, and is such that, for any constant $c > 0$, $L_p p^c \to \infty$ and $L_p p^{-c} \to 0$ as $p \to \infty$. We also define the ‘phase function’ by

$$\rho_\theta(\beta) = \begin{cases} 
\sqrt{(1-\theta)} - \sqrt{(1-\theta)/2 + \frac{1}{2} - \beta)}, & 0 < \beta \leq \frac{1}{2} + (1-\theta)/4, \\
\beta - \frac{1}{2}, & \frac{1}{2} + (1-\theta)/4 < \beta \leq \frac{3}{4}, \\
\{1-\sqrt{(1-\beta)}\}^2, & \frac{3}{4} < \beta < 1.
\end{cases}$$

In the $\beta$–$r$-plane we partition the region $\{\frac{1}{2} < \beta < 1, \rho_\theta(\beta) < r < 1\}$ into three subregions (i), (ii) and (iii) which are defined by $r < \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta) \leq r < \frac{1}{2}$ and $\frac{1}{2} < r < 1$ respectively. The next theorem, which is derived in Delaigle et al. (2010), provides upper and lower bounds for the bootstrap $t$ higher criticism statistic under $H_0$ and $H_1^{(n)}$ respectively.

Theorem 3. Let $p = n^{1/\theta}$, where $\theta \in (0, 1)$ is fixed, and suppose that, for each $1 \leq j \leq p$, the distribution of the respective $X$ satisfies $E(X) = 0$, $E(X^2) = 1$ and $E(|X|^{D_2}) < \infty$, where $D_2$ is chosen so large that equation (2.8) holds with $D_1 > 1/\theta$. Also, take $\alpha_0 = np^{-1} \log(p)$.

(a) Under the null hypothesis $H_0$ in expression (3.1), there is a constant $C > 0$ such that

$$P \{ \text{hc}_n(\alpha_0) \leq C \log(p) \} \to 1 \quad \text{as} \quad n \to \infty.$$

(b) Let $\beta \in \left(\frac{1}{2}, 1\right)$ and $r \in (0, 1)$ be such that $r > \rho_\theta(\beta)$. Under $H_1^{(n)}$ in expression (3.1), where $c^{(j)}$ is modelled as in equations (3.3)–(3.4), we have

$$P \{ \text{hc}_n(\alpha_0) \geq L_p p^{\beta(\beta, r, \theta)} \} \to 1 \quad \text{as} \quad n \to \infty,$$

where
It follows from theorem 3 that, if we set the test to reject the null hypothesis if and only if $\text{hc}_n \geq a_n$, where $a_n/\log(n) \to \infty$ as $n \to \infty$, and $a_n = O(p^\beta)$ where $d < \delta(\beta, r, \theta)$, then, as long as $r > \rho(\beta)$, the probabilities of type I and type II errors tend to 0 as $n \to \infty$ (note that $\delta(\beta, r, \theta) > 0$).

It is also of interest to see what happens when $r < \rho(\beta)$ and below we treat separately the cases $r < \rho(\beta)$ and $\rho(\beta) < r < \rho_0(\beta)$, where $\rho(\beta) \equiv \rho_1(\beta) \geq \rho_0(\beta)$ is the standard phase function that was discussed by Donoho and Jin (2004). See also Fig. 2. We start with the case $r < \rho(\beta)$. There, Ingster (1999) and Donoho and Jin (2004) proved that for the sizes of $\hat{e}_n$ and $\hat{\tau}_n$ that we consider in equations (3.3)–(3.4), even when the underlying distribution of the noise is known to be the standard normal distribution, the sum of the probabilities of type I and type II errors of any test tends to 1 as $n \to \infty$. See also Ingster (2001). Since our testing problem is more difficult than this (in our case the underlying distribution of the noise is estimated from data), in this context also, asymptotically, any test fails if $r < \rho(\beta)$.

It remains to consider the case $\rho(\beta) > r > \rho_0(\beta)$. In the Gaussian model, i.e. when the underlying distribution of the noise is known to be standard normal, it was proved by Donoho and Jin (2004) that there is a higher criticism test for which the sum of the probabilities of type I and type II errors tends to 0 as $n \to \infty$. However, our study does not permit us to conclude that bootstrap $t$ higher criticism will yield a successful test.

There are at least two reasons for possible failure of higher criticism here: first, the sample size $n$ is relatively small and, secondly, we do not have full knowledge of the underlying distribution of background noise. Recall that, in theorems 2 and 3, $P_{\text{F.T.}}(T_0 > \hat{t}_n) = \alpha(1 + a_n)$ for a small error term $a_n$. Ideally, if $a_n = o(p^{-1/2})$ uniformly in $\alpha \in (0, 1)$, the interval between $r = \rho(\beta)$ and $r = \rho_0(\beta)$ vanishes. However, in the present case $n = p^\theta$ where $\theta < 1$. Without further knowledge of the underlying distribution, theorems 2 and 3 suggest that the smallest $a_n$ is $a_n = L_n n^{-1/2} = L_n p^{-\theta/2}$, where $L_n$ lies between two powers of $\log(n)$; this $a_n$ is much larger than $o(p^{-1/2})$.

---

**Fig. 2.** (a) $r = \rho(\beta)$ (Gaussian) (-----), and $r = \rho_0(\beta)$ with $\theta = 0.25$ (-----), $\theta = 0.5$ (-----), and $\theta = 0.75$ (-----) (for each $\theta$, in the region sandwiched by two curves $r = \rho(\beta)$ and $r = \rho_0(\beta)$, higher criticism is successful in the Gaussian case, but maybe not so much in the non-Gaussian case) and (b) magnification of the lower left-hand portion of the graph: the horizontal and vertical axes depict $\beta$ and $r$ respectively.
A potential third reason for difficulties is that, in the idealized Gaussian case (Donoho and Jin, 2004), the success of higher criticism lies in its adaptivity to different signal strengths. When the signal is relatively strong the most informative part of the data is in the tails, but when the signal is weak most of the information is in the centre. Now, the natural scale standardization for the higher criticism statistic is, superficially, \(\alpha(1-\alpha)^{1/2}\) (see equation (3.2)), yet results such as equations (2.8) and (2.9) imply that the bootstrap gives a good approximation to probabilities at the scale \(\alpha\) (in the upper tail) and \(1-\alpha\) (in the lower tail). It follows that the bootstrap approximation for values of \(\alpha\) that are not close to 0 or 1, i.e. which are towards the centre of the distribution, is relatively inaccurate, and this can lead to failure of higher criticism.

It is not difficult to see that, when the underlying distribution is unknown, working under the assumption that it is Gaussian, and directly applying methodology for standard higher criticism, can give poor results. However, when the underlying distribution is Gaussian but we use the bootstrap, performance can also be relatively poor. In this case the bootstrap is encumbered by errors of order \(n^{-1/2}\) that result from estimating skewness, kurtosis and all other cumulants, which in the Gaussian case we know are zero. This renders the bootstrap relatively uncompetitive, although performance is often still reasonable.

The case where \(p\) is exponentially large (i.e. \(n = \log(p)^a\) for some constant \(a > 0\)) can be interpreted as the case \(\theta = 0\), where \(\rho_0(\beta)\) reduces to \(\{1-\sqrt{(1-\beta)}\}^2\). In this case, if \(r > \{1-\sqrt{(1-\beta)}\}^2\) then the sum of probabilities of type I and type II errors of \(hc_n \rightarrow 0\) as \(n \rightarrow \infty\). The proof is similar to that of theorem 3 so we omit it.

4. Numerical properties

First we give numerical illustrations of the results in Section 2.1. In Fig. 3 we compare the right-hand tail of the cumulative distribution functions of \(Z_0\) and \(T_0\) with the right-hand tail of \(\Phi\), denoting the standard normal distribution function, when \(U\) has increasingly heavy tails. We take \(X = \{U - E(U)\}/\text{var}(U)^{1/2}\) where \(U = N |N|\) (moderate tails) or \(N^5 |N|\) (heavier tails), with \(N \sim N(0, 1)\). Fig. 3 shows that \(\Phi\) approximates the distribution of \(T_0\) better than it approximates that of \(Z_0\), and that the approximation of the normal distribution of \(Z_0\) degrades as the distribution of \(X\) becomes more heavy tailed. Fig. 3 also compares the right-hand tail of the inverse cumulative distribution functions, which shows that the normal approximation is more accurate in the tails for \(T_0\) than for \(Z_0\). Unsurprisingly, as the sample size increases the normal approximation for both \(T_0\) and \(Z_0\) becomes more accurate.

Next we illustrate the results in Section 2.2. There we showed that, although \(T_0\) is more robust than \(Z_0\) against heavy tailedness of the distribution \(F_X\) of \(X\), the distribution of \(T_0\) is somewhat more affected by the skewness of \(F_X\). To illustrate the success of the bootstrap in correcting this problem we compare the bootstrap and normal approximations for several skewed and heavy-tailed distributions. In particular, Fig. 4 shows results obtained when \(X = \{U - E(U)\}/\text{var}(U)^{1/2}\), with \(U \sim F(5, 5)\). Since, later in this section, we shall be more interested in approximating quantiles of the distribution of \(T_0\), rather than the distribution itself, then in Fig. 4 we show the right-hand tail of the inverse cumulative distribution function of \(T_0\) and 200 bootstrap estimators of this tail obtained from 200 samples of sizes \(n = 50, n = 100\) or \(n = 250\) simulated from \(F_X\). We also show the inverse cumulative distribution function of the standard normal distribution. Fig. 4 demonstrates clearly that the bootstrap approximation to the tail is more accurate than the normal approximation, and that the approximation improves as the sample size increases. We experimented with other skewed and heavy-tailed distributions, such as other \(F\)-distributions and several Pareto distributions, and reached similar conclusions.
Fig. 3. (a)–(d) Distribution function $F$ and (e)–(h) inverse distribution function $F^{-1}$ of $T_0$ when $U \sim N(1, 1)$ and of an $N(0, 1)$ distribution when $U \sim N(0, 1)$. When (a), (e) $n = 50$, (b), (f) $n = 100$, (c), (g) $U = N(0, 1)$ with $n = 50$ and (d), (h) $U \sim N(0, 1)$.
When implementing the bootstrap, the number \( B \) of bootstrap samples must be taken sufficiently large to obtain reasonably accurate estimators of the tails of the distribution. In general, the larger \( B \), the more accurate the bootstrap approximation is, but in practice we are limited by the capacity of the computer. To obtain a reasonable approximation of the tail up to the quantile \( t_\alpha \), where \( \alpha < \frac{1}{2} \), we found that we should take \( B \) no less than \( \alpha^{-1} \).

Let \( h_c \) and \( h_{c\text{norm}} \) denote respectively the theoretical and the normal versions of the higher criticism statistic, which is defined by the formula at the right-hand side of expression (3.2), replacing there the bootstrap quantiles \( t_\alpha^{(j)} \) by \( t_\alpha^{(j)} \) and \( z_\alpha \) respectively, where \( t_\alpha^{(j)} \) denote the \( 1 - \alpha \) theoretical quantiles of \( T_\alpha^{(j)} \) and \( z_\alpha \) denote the \( (1 - \alpha) \)-quantile of the standard normal distribution. To illustrate the success of the bootstrap in applications of the higher criticism statistic, in our simulations we compared the statistic \( h_c \) which we could use if we knew the distribution \( F_X \), the bootstrap statistic \( h_{c\text{bootstrap}} \) that is defined at expression (3.2), where the unknown quantiles \( t_\alpha^{(j)} \) are estimated as the bootstrap quantities \( t_\alpha^{(j)} \) as discussed in the previous paragraph, and the normal version \( h_{c\text{norm}} \). We constructed histograms of these three versions of the higher criticism statistic, which was obtained from 1000 simulated values calculated under \( H_0 \) or an alternative hypothesis. For any of the three versions, to obtain the 1000 values we generated 1000 samples
Fig. 5. Histograms of hc-statistics under (a)–(d), (i)–(l) and (q)–(t) $H_0$ or under (e)–(h), (m)–(p) and (u)–(x) $H(n)$ when the $X_j$s are standardized $F(5.5)$ variables, $n = 100$, $\theta = \frac{1}{2}$, $p = n^{1/\theta}$, $\varepsilon n = n^{-\beta/\theta}$, and $\tau = 2^{r} \log(p)$, where $r = r_0(\beta)$ and $\beta$ is for the theoretical $hc$, (i)–(t) and (u)–(x) are for $hcn_{\text{norm}}$: (a), (e), (i), (m), (q), (u) $\beta = \frac{1}{2}$; (b), (f), (j), (n), (r), (v) $\beta = \frac{1}{2} + \frac{1}{4}(1 - \theta)$; (c), (g), (k), (o), (s), (w) $\beta = \frac{3}{4}$; (d), (h), (l), (p), (t), (x) $\beta = 1$. 

\[
\begin{align*}
(a) & \quad (b) & \quad (c) & \quad (d) \\
(e) & \quad (f) & \quad (g) & \quad (h) \\
(i) & \quad (j) & \quad (k) & \quad (l) \\
(m) & \quad (n) & \quad (o) & \quad (p) \\
(q) & \quad (r) & \quad (s) & \quad (t) \\
(u) & \quad (v) & \quad (w) & \quad (x)
\end{align*}
\]
of size $n$, of $p$-vectors $(X_1, \ldots, X_p)$. We did this under $H_0$, where the mean of each $X_j$ was 0, and under various alternatives $H_1^{(n)}$, where we set a fraction $\varepsilon_n$ of these means equal to $\tau_n n^{-1/2}$, with $\tau_n > 0$. As in Section 3 we took $p = n^{1/\theta}$, $\varepsilon_n = n^{-\beta/\theta}$ and $\tau_n = \sqrt{2\log(p)}$, where we chose $\beta$ and $\tau$ to be on the frontier of the $r > \rho_0(\beta)$.

Fig. 5 shows the histograms under $H_0$ and under various alternatives $H_1^{(n)}$ located on the frontier ($r = \rho_0(\beta)$, for $\beta = \frac{1}{2}$, $\beta = \frac{1}{2} + \frac{1}{4}(1 - \theta)$, $\beta = \frac{3}{4}$ and $\beta = 1$), when the $X_j$s are standardized $F(5, 5)$ variables, $n = 100$ and $\theta = \frac{5}{7}$. We can see that the histogram approximations to the density of the bootstrap $hc_n$ are relatively close to the histogram approximations to the density of $hc$. By contrast, the histograms in the case of $hc$ norm show that the distribution of $hc$ norm is a poor approximation to the distribution of $hc$, reflecting the inaccuracy of normal quantiles as approximations to the quantiles of heavy-tailed skewed distributions. We also see that, except when $\beta = 1$, the histograms for $hc$ and $hc_n$ under $H_0$ are quite well separated from those under $H_1^{(n)}$. This illustrates the potential success of higher criticism for distinguishing between $H_0$ and $H_1^{(n)}$. By contrast, this property is much less true for $hc$ norm.

We also compared histograms for other heavy-tailed and skewed distributions, such as the Pareto distribution, and reached similar conclusions. Furthermore, we considered skewed but less-heavy-tailed distributions, such as the $\chi^2(10)$ distribution. There also we obtained similar results, but, whereas the bootstrap remained the best approximation, the normal approximation performed better than in heavy-tailed cases. We also considered values of $(\beta, r)$ further from the frontier, and, unsurprisingly since the detection problem became easier, the histograms under $H_1^{(n)}$ became even more separated from those under $H_0$.

Acknowledgements

The work of Aurore Delaigle and Peter Hall was supported by grants and fellowships from the Australian Research Council. We are grateful to Jianqing Fan and Evarist Giné for helpful discussion, and the Associate Editor and a referee for insightful comments.

Appendix A: Technical arguments

A.1. Preliminaries

Let $T_r$ be as in expression (2.11). Then the following result can be proved by using arguments of Wang and Hall (2009).

**Theorem 4.** Let $B > 1$ denote a constant. Then,

$$
\frac{P(T_r > x)}{1 - \Phi(x - c)} = \exp\left\{-\frac{1}{2} n^{-1/2} (2c^2 - 3cx^2 + x^4)\right\} \{1 + \theta(c, n, x)\} \{1 + (1 + |x|)n^{-1/2} + (1 + |x|)^{4n}n^{-1}\} 
$$

(A.1)

as $n \to \infty$, where the function $\theta$ is bounded in absolute value by a finite, positive constant $C_1(B)$ (depending only on $B$), uniformly in all distributions of $X$ for which $E(|X|^4) \leq B$, $E(X^2) = 1$ and $E(X) = 0$, and uniformly in $c$ and $x$ satisfying $0 \leq x \leq Bn^{1/4}$ and $0 \leq c \leq ux$, where $0 < u < 1$.

We shall employ theorem 4 to prove theorem 5 below. Details are given in Delaigle et al. (2010). Take $F$ to be any subset of the class of distributions $F$ of the random variable $X$, such that $E(|X|^{6+\epsilon}) \leq B$ for some $\epsilon > 0$ and a constant $1 < B < \infty$, $E(X) = 0$ and $E(X^2) = 1$. Recall the definition of $T_0^*$ in equation (2.6), let $t = t_0$ and $t = t_\alpha$ denote the respective solutions of $P(T_0^* > t) = \alpha$ and $P(T_0^* > t|X) = \alpha$, and recall that $z_\alpha = (1 - \Phi)^{-1}(\alpha)$. Take $\eta \in (0, \epsilon/4(6 + \epsilon))$, and let $T_r$ and $t_\alpha$ denote independent random variables with the specified marginal distributions.

**Theorem 5.** Let $B > 1$ denote a constant. Then,

$$
P_{T_r}(T_r > t_\alpha) = P_{T_r}(T_r > t_\alpha) \exp\left\{\frac{1}{2} n^{-1/2} c(3z_\alpha^2 - c^2)\right\} \{1 + O\{1 + z_\alpha\}n^{-1/2} + (1 + z_\alpha)^{4n}n^{-1}\} 

+ O\left[\sum_{k=1}^{3} P_{T_r}\left\{\left|\frac{1}{n} \sum_{i=1}^{n} (1 - E)X_i^2\right| > n^{-1/4-\eta}\right\} + P_{T_r}\left\{\left|\frac{1}{n} \sum_{i=1}^{n} (1 - E)X_i^2\right| > B\right\}\right] 
$$

(A.2)
as \( n \to \infty \), uniformly in all \( F \in \mathcal{F} \) and in all \( c \) and \( z_\alpha \) satisfying \( 0 \leq z_\alpha \leq Bn^{1/4} \) and \( 0 \leq c \leq uz_\alpha \), where \( 0 < u < 1 \).

### A.2. Proof of theorem 2

The following theorem can be derived from results of Adamczak (2008).

**Theorem 6.** If \( Y_1, \ldots, Y_n \) are independent and identically distributed random variables with zero mean and unit variance and satisfying

\[
P(|Y| > y) \leq K_1 \exp(-K_2 y^2)
\]

for all \( y > 0 \), where \( K_1, K_2, \xi > 0 \), then for each \( \lambda > 1 \) there are constants \( K_3, K_4 > 0 \), depending only on \( K_1, K_2, \xi \) and \( \lambda \), such that, for all \( y > 0 \),

\[
P\left( \left| \sum_{i=1}^n Y_i \right| > y \right) \leq 2 \exp\left(-\frac{y^2}{2\lambda n}\right) + K_3 \exp\left(-\frac{y^2}{K_4}\right).
\]

We use theorem 6 to bound the remainder terms in theorem 5. If \( P_F(|X| > x) \leq C_1 \exp(-C_2 x^\xi) \) and we take \( Y = (1 - E)X^k \) for an integer \( k \), then inequality (A.3) holds for constants \( K_1, K_2 \) depending on \( C_1, C_2 \) and \( \xi_1 \), and with \( \xi = \xi_1/k \). In particular, for all \( x > 0 \),

\[
P_F\left\{ \sum_{i=1}^n (1 - E)X_i^k \right\} \geq x \var(X^k)^{1/2} \leq 2 \exp\left(-\frac{x^2}{2\lambda n}\right) + K_3 \exp\left(-\frac{x^2}{K_4}\right).
\]

Taking \( k = 1, 2, 3 \), and \( x = x_{t_0} = \text{constant} \times n^{3/4 - \eta} \) for some \( \eta > 0 \) or \( k = 4 \) and \( x = x_{t_0} = \text{constant} \), we deduce that in each of these settings

\[
P_F\left\{ \sum_{i=1}^n (1 - E)X_i^k \right\} \geq x_{t_0} \begin{cases} O\{\exp(-n^{\xi_1/4k - \eta})\} & \text{if } k = 1, 2, 3, \\ O\{\exp(-n^{\xi_1/4 / K_3})\} & \text{if } k = 4, \end{cases}
\]

where \( \eta_2 > 0 \) decreases to 0 as \( \eta_1 \downarrow 0 \). Therefore the \( O[\cdots] \) remainder term in equation (A.2) equals \( O\{\exp(-n^{\xi_1/16 - \eta})\} \), and so theorem 2 is implied by theorem 5.

### A.3. Proof of result (2.12)

Note that, by theorem 5 in Appendix A.1, theorems 1 and 2 continue to hold if we replace the left-hand sides of equations (2.8) and (2.10) by \( P_F(T_c > t_n) \), provided that we also replace the factor \( \alpha \) on the right-hand sides by \( P_F(T_c > t_n) \). The uniformity with which equations (2.8) and equation (2.10) hold now extends (in view of theorem 5) to \( c \) such that \( 0 \leq c \leq uz_\alpha \) with \( 0 < u < 1 \), as well as to \( \alpha \) satisfying \( 0 \leq z_\alpha \leq Bn^{1/4} \).

### References

Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13, 1000–1034.

Almirantis, Y. and Provata, A. (1999) Long- and short-range correlations in genome organization. *J. Statist. Phys.*, 97, 233–262.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57, 289–300.

Bernhard, G., Klein, M. and Hommel, G. (2004) Global and multiple test procedures using ordered p-values—a review. *Statist. Pap.*, 45, 1–14.

Blair, R. C., Troendle, J. F. and Beck, R. W. (1996) Control of familywise errors in multiple endpoint assessments via stepwise permutation tests. *Statist. Med.*, 15, 1107–1121.

Cai, T. and Jin, J. (2010) Optimal rate of convergence of estimating the null density and the proportion of non-null effects in large-scale multiple testing. *Ann. Statist.*, 38, 100–145.

Clarke, S. and Hall, P. (2009) Robustness of multiple testing procedures against dependence. *Ann. Statist.*, 37, 332–358.

David, J.-P., Strodte, C., Vontas, J., Nikou, D., Vaughan, A., Pignatelli, P. M., Louis, P., Hemingway, J. and Ranson, J. (2005) The Anopheles gambiae detoxification chip: a highly specific microarray to study metabolic-based insecticide resistance in malaria vectors. *Proc. Natl. Acad. Sci. USA*, 102, 4080–4084.

Delaigle, A. and Hall, P. (2009) Higher criticism in the context of unknown distribution, non-independence and classification. In *Perspectives in Mathematical Sciences II: Probability and Statistics* (eds N. Sastry, M. Delampady, B. Rajeev and T. S. S. R. K. Rao), pp. 109–138. Singapore: World Scientific Press.
Wu, W. B. (2008) On false discovery control under dependence.

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

Donoho, D. L. and Jin, J. (2006) Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.*, 34, 2980–3018.

Dudouit, S., Shaffer, J. P. and Boldrick, J. C. (2003) Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, 18, 73–103.

Efron, B. (1987) Better bootstrap confidence intervals (with discussion). *J. Am. Statist. Ass.*, 82, 171–200.

Fan, J., Hall, P. and Yao, Q. (2007) To how many simultaneous hypothesis tests can normal, Student’s t or bootstrap calibration be applied? *J. Am. Statist. Ass.*, 102, 1282–1288.

Fan, J. andLv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B*, 70, 849–911.

Finner, H. and Roters, M. (2002) Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.*, 30, 220–238.

Genovese, C. and Wasserman, L. (2004) A stochastic process approach to false discovery control. *Ann. Statist.*, 32, 1035–1061.

Giné, E., Götze, F. and Mason, D. M. (1997) When is the Student t-statistic asymptotically standard normal? *Ann. Probab.*, 25, 1514–1531.

Hall, P. (1988) Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.*, 16, 927–985.

Hall, P. (1990) On the relative performance of bootstrap and Edgeworth approximations of a distribution function. *J. Multiv. Anal.*, 35, 108–129.

Hall, P. and Wang, Q. (2004) Exact convergence rate and leading term in central limit theorem for Student’s t statistic. *Ann. Probab.*, 32, 1419–1437.

Hall, P. and Wang, Q. (2010) Strong approximations of level exceedences related to multiple hypothesis testing. *Bernoulli*, 16, 418–434.

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

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

Jin, J. (2008) Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators (with discussion). *J. R. Statist. Soc. B*, 70, 461–493.

Jin, J. and Cai, T. (2007) Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. Am. Statist. Ass.*, 102, 496–506.

Kesselman, H. J., Cribbie, R. and Holland, B. (2002) Controlling the rate of Type I error over a large set of statistical tests. *Br. J. Math. Statist. Psychol.*, 55, 27–39.

Kulinskaya, E. (2009) On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. *Biometrika*, 96, 201–211.

Lang, T. A. and Secic, M. (1997) *How to Report Statistics in Medicine: Annotated Guidelines for Authors*. Philadelphia: American College of Physicians.

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.

Linnik, Ju. V. (1961) Limit theorems for sums of independent quantities, taking large deviations into account: I. *Teor. Veroj. Primen.*, 7, 145–163.

Mansilla, R., De Castillo, N., Govezensky, T., Miramontes, P., José, M. and Coho, G. (2004) Long-range correlation in the whole human genome. *Universidad Nacional Autonoma de México, México*. (Available from http://arxiv.org/pdf/q-bio/0402043v1.)

Petrov, V. V. (1975) *Sums of Independent Random Variables*. Berlin: Springer.

Pigot, I. (2000) Basic concepts of multiple tests—a survey. *Statist. Pap.*, 41, 3–36.

Sarkar, S. K. (2006) False discovery and false nondiscovery rates in single-step multiple testing procedures. *Ann. Statist.*, 34, 394–415.

Shao, Q.-M. (1999) A Cramér type large deviation result for Student’s t-statistic. *J. Theor. Probab.*, 12, 385–398.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64, 479–498.

Student (1908) The probable error of a mean. *Biometrika*, 6, 1–25

Takada, T., Hasegawa, T., Ogura, H., Tanaka, M., Yamada, H., Komura, H. and Ishii, Y. (2001) Statistical filter for multiple test noise on fMRI. *Syst. Comput. Jpn.*, 32, 16–24.

Tsamane, A. C. and Dunnett, C. W. (1999) Stepwise multiple test procedures with biometric applications. *J. Statist. Planng Inf.*, 82, 55–68.

Wang, Q. and Hall, P. (2009) Relative errors in central limit theorems for Student’s t statistic, with applications. *Statist. Sin.*, 19, 343–354.

Wu, W. B. (2008) On false discovery control under dependence. *Ann. Statist.*, 36, 364–380.