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

Aurore Delaigle[1] Peter Hall[1,2] and Jiashun Jin[3]

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

[2] Department of Statistics, University of California at Davis, Davis, CA 95616, USA.

[3] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract: 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, for example robustness against heavy tailed sampling distributions, that were not explicitly considered until relatively recently. In this paper 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, for example 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.

AMS SUBJECT CLASSIFICATION. Primary 62G32; Secondary 62F30, 62G90, 62H15, 62H30

Short title. Student's t statistic

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 of the latter 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. In order to standardise 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 Studentised 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 Studentising have been discovered. In particular, the t statistic's high degree of robustness against heavy-tailed data has been quantified. For example, Giné, Götze and Mason (1997) have shown that a necessary and sufficient condition for the Studentised mean to have a limiting standard normal distribution is that the sampled distribution lie 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 Studentised mean to normality is strictly faster than that for the conventional mean, normalised 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 Studentised 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 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, 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 ones. 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 Studentised 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 (that is, 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), Cai and Jin (2010) and Kulinskaya (2009). The same issues arise in the case of methods for signal detection, for example those based on Student's t versions of higher criticism; see Donoho and Jin (2004), Jin (2007) and Delaigle and Hall (2009). Work in the context of multiple hypothesis testing includes that of Lang and Secic (1997, p. 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 Studentised ratio in a variety of high dimensional settings. Our results reveal several advantages of Studentising. 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 Studentised form.

Section 2.1 draws together several known results in the literature in order to demonstrate the robustness of the t ratio in the context of high level exceedences. 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 exceedence is either polynomially or exponentially small. Section 2.4 shows how these properties can be applied to high dimensional problems, involving potential exceedences 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 drawn in sections 2 and 3 are illustrated numerically in section 4, the underpinning theoretical arguments are summarised in section 5, and detailed arguments are given by Delaigle *et al.* (2010).

2 Main conclusions and theoretical properties

2.1 Advantages and drawbacks of studentising 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 we want to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Two common test statistics for this problem are the standardised mean Z_0 and the Studentised mean T_0, defined by $Z_0 = n^{1/2}\,\bar{X}$ and $T_0 = Z_0/S$ where

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,, \quad S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{2.1}$$

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

In practice, experience with the context often suggests the standardisation 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 impact of skewness. However, we shall show that, compared to 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

$$E|X|^3 < \infty \tag{2.2}$$

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 values of x that satisfy $0 < x \le \epsilon\, n^{1/6}$, for any positive sequence ϵ that converges to zero (Shao, 1999). This level of accuracy applies also to the normal approximation to the distribution of the nonstudentised mean, $\bar{X}$, except that we must impose a condition much more severe than (2.2). In particular, $P(Z_0 > x)/\{1 - \Phi(x)\} \to 1$, uniformly in $0 < x \le n^{(1/6)-\eta}$, for each fixed $\eta > 0$, if and only if

$$E\big\{ \exp\big(|X|^c\big)\big\} < \infty \quad \text{for all} \quad c \in \big(0, \tfrac{1}{2}\big)\,; \tag{2.3}$$

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 (2.2).

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

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

$$P(Z_0 > x) = \{1 - \Phi(x)\}\, \big\{1 + n^{-1/2}\, \tfrac{1}{6}\, x^3\, \gamma + o\big(n^{-1/2}\, x^3\big)\big\}\,, \tag{2.5}$$

uniformly in x satisfying $\lambda_n \le x \le n^{1/6}\lambda_n$, for a sequence $\lambda_n \to \infty$, and where Φ is the standard normal distribution function and $\gamma = E(X^3)$ (Shao, 1999; Petrov, 1975, Chap. 8). (Property (2.2) is sufficient for (2.4) if $x \to \infty$ and $n^{-1/2}\, x^3 \to 0$ as $n \to \infty$, and (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 $\gamma \ne 0$ 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 $\mathrm{N}(0, \sigma^2)$, and therefore relates to the case of zero skewness. Taking $\gamma = 0$ in (2.4) we see that, when T_0 has Student's t distribution with n or $n-1$ degrees of freedom, we have $P(T_0 > x) = \{1-\Phi(x)\} \big\{1+o\big(n^{-1/2} x^3\big)\big\}$. It can be deduced that the results derived in (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's 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's 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 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 (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 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 a normal approximation. However, as for the normal approximation, the latter 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 $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ denote a resample drawn by sampling randomly, with replacement, from $\mathcal{X} = \{X_1, \ldots, X_n\}$, and put

$$\bar{X}^* = \frac{1}{n} \sum_{i=1}^{n} X_i^* , \quad S^{*2} = \frac{1}{n} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)^2 , \quad T_0^* = n^{1/2} (\bar{X}^* - \bar{X})/S^* . \tag{2.6}$$

The bootstrap approximation to the distribution function $G(t) = P(T_0 \le t)$ is $\widehat{G}(t) = P(T_0^* \le t \,|\, \mathcal{X})$, and the bootstrap approximation to the quantile $t_\alpha = (1-G)^{-1}(\alpha)$ is

$$\widehat{t}_\alpha = \big(1 - \widehat{G}\big)^{-1}(\alpha) . \tag{2.7}$$

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, defined below (2.6), can be written as $G(t) = H(t \,|\, F)$, where $H(\,\cdot\,|\, F)$ is a functional. In this notation the bootstrap distribution $\widehat{G}$ is just $H(\,\cdot\,|\,\widehat{F})$, where $\widehat{F}$ 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 $\widehat{F}$, is a consistent estimator of the skewness of the distribution of X_i, and so skewness can be expected to be captured accurately by $\widehat{G}$ because it is inherited from $\widehat{F}$.

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 $\widehat{t}_\alpha$ of the bootstrap test is random and constructed from the sample then, to avoid confusion, we should arguably use different notations for T_0 and the generic variable. However, to simplify notation we keep using T_0 for a generic random variable distributed like T_0. This means that we write the level of the test as $P(T_0 > \widehat{t}_\alpha)$, but here T_0 denotes a generic random variable independent of the sample, whereas $\widehat{t}_\alpha$ denotes the random variable defined at (2.7) and calculated from the sample. In particular, here T_0 is independent of $\widehat{t}_\alpha$.

Define $z_\alpha = (1-\Phi)^{-1}(\alpha)$, and write P_F for the probability measure when $\mathcal{X}$ 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*

$$P_F(T_0 > \widehat{t}_\alpha) = \alpha \left[1 + O\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\right] + O\big(n^{-D_1}\big) \qquad (2.8)$$

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

The theorem can be deduced by taking $c = 0$ in Theorem B in section 5.1, 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 (2.8), the following result, which can be deduced from Theorem A in section 5.1 for each $B > 1$ such that $E|X|^4 < B$ and $0 \le z_\alpha \le B\, n^{1/4}$:

$$P_F(T_0 > z_\alpha) = \alpha \,\exp\big(- n^{-1/2}\, \tfrac{1}{3}\, z_\alpha^3\, \gamma\big) \left[1 + O\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\right]. \qquad (2.9)$$

Comparing (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 (2.8) can be dropped if the distribution of X is assumed to be sufficiently smooth. The version of (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 (2.9), since neither the normal approximation implicit in the use of z_α on the left-hand side of (2.9), nor the factor $\exp(-n^{-1/2}\, \frac{1}{3}\, \gamma)$ on the right-hand side, adequately compensates for skewness.

To appreciate that the $1 + O\{\ldots\}$ remainder term in (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 $\widehat{t}_\alpha$ by a factor $1 + O(n^{-1/2}\, z_\alpha^{-1} + n^{-1}\, z_\alpha^2)$. However, since $\widehat{t}_\alpha = z_\alpha\, [1 - n^{-1/2}\, \frac{1}{3}\, \{1 + o_p(1)\}\, z_\alpha^2\, \gamma]$, where the skewness correction $-n^{-1/2}\, \frac{1}{3}\, \{1 + o_p(1)\}\, z_\alpha^2\, \gamma$ is of strictly larger order than $n^{-1/2}\, z_\alpha^{-1} + n^{-1}\, z_\alpha^2$, then an adjustment for the remainder in (2.8) would be minor relative to the skewness correction. In this sense the remainder term in (2.8), and in related formulae below, has a relatively minor impact 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 applied to Student's t statistic can have a significant, positive effect on our capacity for avoiding false discoveries using methods such as those of Benjamini and Hochberg (1995), Blair *et al.* (1996), Storey (2002) and Genovese and Wasserman (2006).

The size of the $O(n^{-D_1})$ remainder in (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 α of size p^{-1}, in the sense that $p\alpha$ is bounded away from zero and infinity 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 "$P_F(T_0 > \widehat{t}_\alpha)/\alpha = 1 + o(1)$ for $p = o(n^{D_1})$ and uniformly in the class of distributions F of X for which $E(X) = 0$ and $E(|X|^{D_2}) \le B\,(EX^2)^{D_2/2}$."

On the other hand, 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 the theorem is given in section 5.2.

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*

$$P_F(T_0 > \widehat{t}_\alpha) = \alpha\left[1 + O\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\right] + O\big\{\exp\big(-n^{D_1}\big)\big\} \tag{2.10}$$

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

Theorem 2 allows us to repeat all the remarks 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 supported on a compact domain. Note that our condition is still less restrictive than assuming that the distribution is normal, as is done in many papers treating high dimensional problems, such as for example Fan and Lv (2008).

2.3 Effect of a nonzero 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 Studentised mean rather than the standardised 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 Studentised 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 notationally simpler if we continue to assume that $E(X) = 0$ and $\mathrm{var}\,(X) = 1$, 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 zero. We define

$$Z_c = n^{1/2}\left(\bar{X} + c\,n^{-1/2}\right), \quad T_c = Z_c/S\,. \tag{2.11}$$

Here we take μ 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 zero. 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 $P_F(T_c > \widehat{t}_\alpha)$, where $c > 0$ and $\widehat{t}_\alpha$ is given by (2.7), and when T_c and $\widehat{t}_\alpha$ are computed from independent data. We claim that in this setting the results discussed in section 2.2 continue to hold. In particular, versions of (2.8) and (2.10) in the present setting are:

$$P_F(T_c > \widehat{t}_\alpha) = P_F(T_c > t_\alpha)\,\Big[1 + O\big\{(1+z_\alpha)\,n^{-1/2} + (1+z_\alpha)^4\,n^{-1}\big\}\Big] + R\,, \tag{2.12}$$

where the remainder term R has either the form in (2.8) or that in (2.10), depending on whether we assume existence of polynomial or exponential moments, respectively. In particular, if we take $R = O(n^{-D_1})$ then (2.12) holds uniformly in all distributions F of the random variable X such that $E(|X|^{D_2}) \le B\,(EX^2)^{D_2/2}$ and $E(X) = 0$, and in all α satisfying $0 \le z_\alpha \le B\,n^{1/4}$, provided that D_2 is sufficiently large; and in the same sense, but with $R = O\{\exp(-n^{D_1})\}$ where $D_1 \in (0, \frac{3}{8}]$, (2.12) holds if we replace the assumption $E(|X|^{D_2}) \le B\,(EX^2)^{D_2/2}$ by $P\{|X| > x/(EX^2)^{1/2}\} \le C\,\exp(-x^{D_2})$, 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 section 5.3. 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 (2.8) and (2.10)).

An alternative form of (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 α. This can be done if we note that, in view of Theorem A in section 5.1,

$$\begin{aligned} P_F(T_c > t_\alpha) &= \{1 - \Phi(t_\alpha)\}\,\exp\Big\{ - n^{-1/2}\,\tfrac{1}{6}\,\big(2\,t_\alpha^3 - 3\,c\,t_\alpha^2 + c^3\big)\,\gamma\Big\}\,\frac{1 - \Phi(t_\alpha - c)}{1 - \Phi(t_\alpha)} \\ &\qquad \times \Big[1 + \theta(c,n,t_\alpha)\,\Big\{(1 + t_\alpha)\,n^{-1/2} + (1 + t_\alpha)^4\,n^{-1}\Big\}\Big] \\ &= \alpha\,\exp\big\{n^{-1/2}\,\tfrac{1}{6}\,c\,\big(3\,t_\alpha^2 - c^2\big)\,\gamma\big\}\,\frac{1 - \Phi(t_\alpha - c)}{1 - \Phi(t_\alpha)} \\ &\qquad \times \Big[1 + \theta_1(c,n,t_\alpha)\,\Big\{(1 + t_\alpha)\,n^{-1/2} + (1 + t_\alpha)^4\,n^{-1}\Big\}\Big], \end{aligned} \tag{2.13}$$

where γ denotes skewness, θ_1 has the same interpretation as θ in Theorem A, and the last identity follows from the definition of t_α. Combining this property with (2.12) it can be shown that

$$\begin{aligned} P_F(T_c > \widehat{t}_\alpha) &= \alpha\,\exp\big\{n^{-1/2}\,\tfrac{1}{6}\,c\,\big(3\,t_\alpha^2 - c^2\big)\,\gamma\big\}\,\frac{1 - \Phi(t_\alpha - c)}{1 - \Phi(t_\alpha)} \\ &\qquad \times \Big[1 + O\big\{(1 + z_\alpha)\,n^{-1/2} + (1 + z_\alpha)^4\,n^{-1}\big\}\Big] + R, \end{aligned} \tag{2.14}$$

where R satisfies the properties given below (2.12).

2.4 Relationships among many events $T_c > \widehat{t}_\alpha$

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; that is, we observe $T^{(j)}_{c^{(j)}}$ for $1 \leq j \leq p$, where $c^{(1)}, \ldots, c^{(p)}$ are constants and the random variables $T^{(j)}_{c^{(j)}}$ are, for different values j, computed from independent datasets. We assume that $T^{(j)}_{c^{(j)}}$ is defined as at (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, treated in Theorems 1 and 2, it is important that the t-statistic $T^{(j)}_{c^{(j)}}$ and the corresponding quantile estimator $\widehat{t}^{(j)}_{\alpha}$ be independent for each j. However, as noted in section 2.2, this is not a problem since $T^{(j)}_{c^{(j)}}$ represents a generic random variable, and only $\widehat{t}^{(j)}_{\alpha}$ is calculated from the sample.

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

Cases where the statistics are computed from weakly dependent data can be addressed using results of Hall and Wang (2010). That work treats instances where the variables $T^{(j)}_{c^{(j)}}$ are computed from the first n components in respective data streams $\mathcal{S}_j = (X_{j1}, X_{j2}, \ldots)$, 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 standardise scale, and in particular places the means $c^{(j)}$ on the same scale for each j.) Assuming this is the case, we shall suppose too 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 $\mathcal{S}_1, \mathcal{S}_2, \ldots$ are k-dependent for some fixed $k \geq 1$, (c) z_α is bounded between two constant multiples of $(\log p)^{1/2}$, (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 realisations that arise with probability no

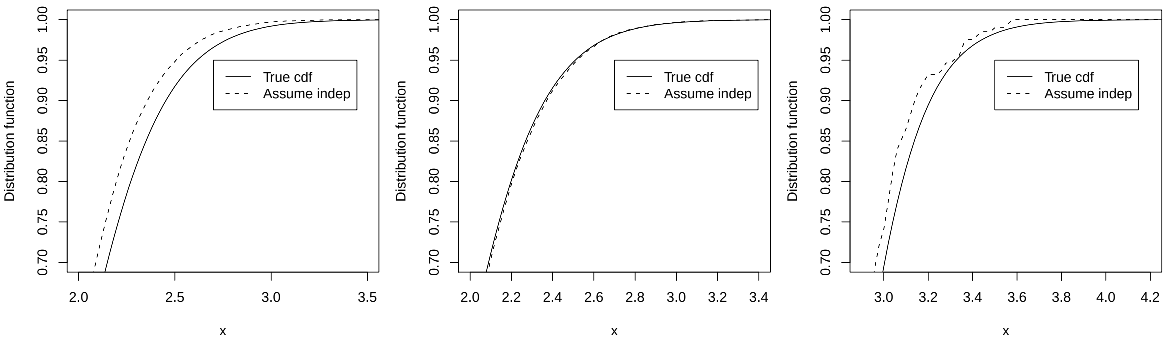

Figure 1: Comparison of the joint distribution function of $(T_0^{(1)}, \ldots, T_0^{(p)})$ (denoted by "True cdf") with the product of the distributions of the univariate components $T_0^{(k)}$, $k = 1, \ldots, p$ (denoted by "Assume indep"), when $\epsilon_k \sim$ standardised Pareto(5,5), $n = 50$ and, from left to right, $(p, \theta) = (100, 0.5)$, $(p, \theta) = (100, 0.2)$, $(p, \theta) = (10000, 0.2)$. 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.

greater than $1 - O\{p \exp(-C\, z_\alpha^2)\}$, where $C > 0$; the t-statistics $T_{c(j)}^{(j)}$ can be considered to be independent. In particular, it can be stated that with probability $1 - O\{p \exp(-C\, z_\alpha^2)\}$ there are no clusters of level exceedences caused by dependence among 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 $\mathcal{S}_{rj+1} = \ldots = \mathcal{S}_{rj+k}$ 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 $S_j^2 = n^{-1} \sum_{i \leq n} (X_{ji} - \bar{X}_j)^2$ is the empirical variance computed from the first n data in the stream $\mathcal{S}_j$. 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.

To illustrate these properties we calculated the joint distribution of $(T_0^{(1)}, \ldots, T_0^{(p)})$ 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 $T_0^{(k)}$, $k = 1, \ldots, p$.

For $k = 1, \ldots, p$ we took $X_k = (U_k - EU_k)/\sqrt{\operatorname{var} U_k}$ 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 i.i.d. random variables. Figure 1 depicts the resulting distribution functions for several values of θ and p, when the sample size n was 50 and the ϵ_js were from a standardised 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 θ (and thus the dependence) increases. The figure 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 show that the advantages established in section 2 for bootstrap t methods carry over to sparse signal detection.

Assume we observe $X_{1j}, \ldots, X_{nj}$, for $1 \le j \le p$, 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} if often used to represent the log-intensity associated with the ith subject and the jth gene, μ_j represents the mean expression level associated with the jth 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 among components. Let $E(X_{1j}) = c^{(j)}$. The problem of signal detection is to test

$$H_0\colon \text{ all } c^{(j)}\text{s are zero, against } \quad H_1^{(n)}\colon \text{ a small fraction of the } c^{(j)}\text{s is nonzero.} \quad (3.1)$$

For simplicity, in this section we assume that each $c^{(j)} \ge 0$, but a similar treatment can be given where nonzero $c^{(j)}$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 centered data have a known distribution, non-Studentised means were used, and the bootstrap was not employed.) As in section 2.4, let $T^{(j)}_{c^{(j)}}$ be the Studentised statistic for the jth component, and let $\widehat{t}^{(j)}_{\alpha}$ be the bootstrap estimator of the $1-\alpha$ quantile of the distribution of $T^{(j)}_0$, both calculated from the data $X_{1j},\ldots X_{nj}$. We suggest the following bootstrap t higher criticism statistic:

$$\mathrm{hc}_n(\alpha_0) = \max_{\alpha=i/p,\,1\le i\le \alpha_0 p} \{p\,\alpha\,(1-\alpha)\}^{-1/2} \sum_{j=1}^{p} \bigl\{I\bigl(T^{(j)}_{c^{(j)}} > \widehat{t}^{(j)}_{\alpha}\bigr) - \alpha\bigr\}\,, \tag{3.2}$$

where $\alpha_0 \in (0,1)$ is small enough for the statistic hc_n at (3.2) to depend only on indices j for which $T^{(j)}_{c^{(j)}}$ is relatively large. This exploits the excellent performance of bootstrap approximation to the distribution of the Studentised 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 $\mathrm{hc}_n(\alpha_0)$ 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 to normal quantiles, established in section 2, suggest that in the higher criticism context, too, 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 (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 zero 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 nonzero. It is standard to take $c^{(j)} = 0$ for all but a fraction ϵ_n of js, 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

$$\epsilon_n = p^{-\beta} = n^{-\beta/\theta}, \tag{3.3}$$

where $\beta \in (0,1)$ is a fixed parameter. Among these values of β the range $0 < \beta < \frac{1}{2}$ is the least interesting, because there the proportion of nonzero signals is so high that it is possible to estimate the signal with reasonable accuracy, rather than just determine its existence. See Donoho and Jin (2004). Therefore we focus on the most interesting range, which is $\frac{1}{2} < \beta < 1$. For $\beta \in (\frac{1}{2}, 1)$ the most interesting values of τ_n are $\tau_n \asymp \sqrt{2\log p}$, with $\tau_n < \sqrt{2\log p}$. Taking $\tau_n = o(\sqrt{2\log p})$ 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 nonzero are of the same size as, or larger than, the largest values of the signal-free $T^{(j)}_{c^{(j)}}$s. In light of this we consider nonzero means of size

$$\tau_n = \sqrt{2r\log p} = \sqrt{2(r/\theta)\log n}\ , \tag{3.4}$$

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

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

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

In the β-r plane we partition the region $\{\frac{1}{2} < \beta < 1, \rho_\theta(\beta) < r < 1\}$ into three subregions (i), (ii), and (iii) defined by $r < \frac{1}{4}(1-\theta)$, $\frac{1}{4}(1-\theta) \leq r < \frac{1}{4}$, and $\frac{1}{4} < r < 1$, respectively. The next theorem, derived in the longer version of this paper (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* (2.8) *holds with $D_1 > 1/\theta$. Also, take $\alpha_0 = n\,p^{-1}\log p$. Then*
(a) Under the null hypothesis H_0 in (3.1)*, there is a constant $C > 0$ such that*

$$P\big\{\mathrm{hc}_n(\alpha_0) \leq C\log p\big\} \to 1 \ \textit{as}\ n \to \infty\,.$$

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

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

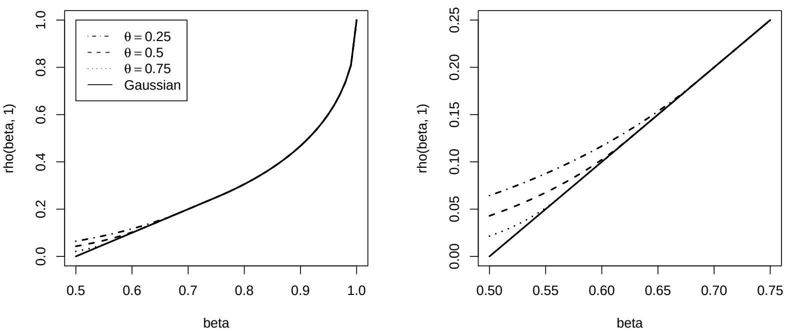

Figure 2: Left: $r = \rho(\beta)$ (black) and $r = \rho_\theta(\beta)$ with $\theta = 0.25$ (blue), 0.5 (green), and 0.75 (red). For each θ, in the region sandwiched by two curves $r = \rho(\beta)$ and $r = \rho_\theta(\beta)$, higher criticism is successful in the Gaussian case, but maybe not so much in the non-Gaussian case. Right: magnification of lower left portion of graph. The horizontal and vertical axes depict β and r, respectively.

where

$$\delta(\beta, r, \theta) = \begin{cases} \frac{1}{2} - \beta + (1-\theta)/2 - (\sqrt{(1-\theta)} - \sqrt{r})^2, & \text{if } (\beta, r) \text{ is in region (i)}, \\ r - \beta + \frac{1}{2}, & \text{if } (\beta, r) \text{ is in region (ii)}, \\ 1 - \beta - (1 - \sqrt{r})^2, & \text{if } (\beta, r) \text{ is in region (iii)}. \end{cases}$$

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

It is also of interest to see what happens when $r < \rho_\theta(\beta)$, and below we treat separately the cases $r < \rho(\beta)$ and $\rho(\beta) < r < \rho_\theta(\beta)$, where $\rho(\beta) \equiv \rho_1(\beta) \geq \rho_\theta(\beta)$ is the standard phase function discussed by Donoho and Jin (2004). We start with the case $r < \rho(\beta)$. There, Ingster (1999) and Donoho and Jin (2004) proved that for the sizes of ϵ_n and τ_n that we consider in (3.3)–(3.4), even when the underlying distribution of the noise is known to be the standard normal, 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 too, asymptotically, any test fails if $r < \rho(\beta)$.

It remains to consider the case $\rho(\beta) < r < \rho_\theta(\beta)$. 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 2–3, $P_F(T_0 > \widehat{t}_\alpha) = \alpha\,(1 + a_n)$ 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 $r = \rho_\theta(\beta)$ vanishes. However, in the present case $n = p^\theta$ where $\theta < 1$. Without further knowledge of the underlying distribution, Theorems 2–3 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 $o(p^{-1/2})$.

A potential third reason for difficulties is that, in the idealised 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 tail, but when the signal is weak most of the information is in the centre. Now, the natural scale standardisation for the higher criticism statistic is, superficially, $\{\alpha\,(1-\alpha)\}^{1/2}$ (see (3.2)), yet results such as (2.8) and (2.9) imply that the bootstrap gives a good approximation to probabilities at the scale α (in the upper tail) and $1-\alpha$ (in the lower tail). It follows that the bootstrap approximation for values of α 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. On the other hand, 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 $\rho_\theta(\beta)$ 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 tends to 0 as n tends to ∞. The proof is similar to that of Theorem 3 so we

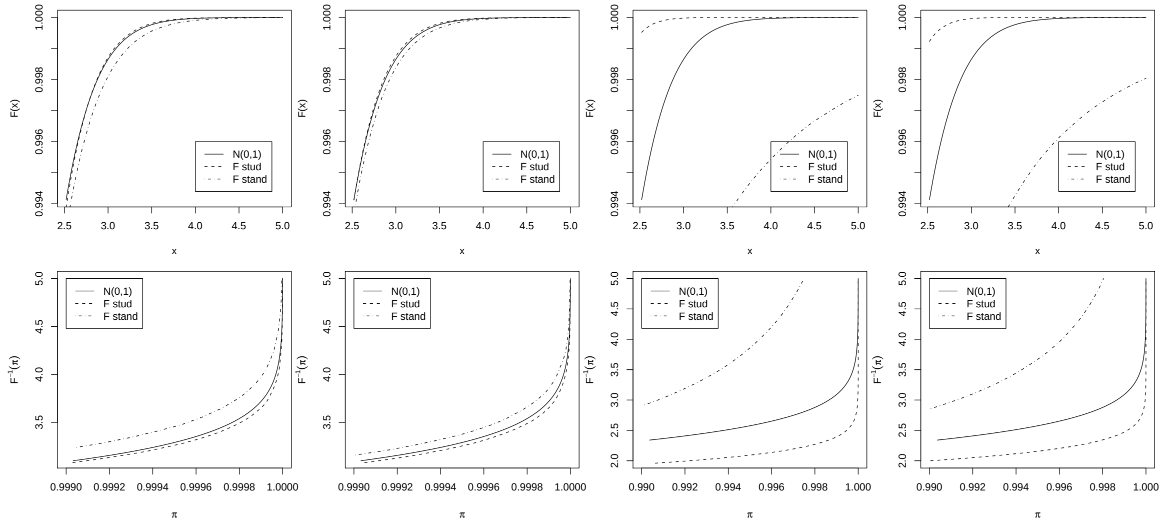

Figure 3: Distribution function (F), top, and inverse distribution function (F^{-1}), bottom, of T_0 (F stud), Z_0 (F stand) and of a N(0,1) when $U = N|N|$, when, from left to right $n = 50$, $U = N|N|$ with $n = 100$, $U = N^5|N|$ with $n = 50$, $U = N^5|N|$ with $n = 100$ and where $N \sim \mathrm{N}(0,1)$.

omit it.

4 Numerical properties

First we give numerical illustrations of the results in section 2.1. In Figure 3 we compare the right tail of the cumulative distribution functions of Z_0 and T_0 with the right tail of Φ, denoting the standard normal distribution function, when U has increasingly heavy tails. We take $X = (U - EU)/(\mathrm{var}\, U)^{1/2}$ where $U = N|N|$ (moderate tails) or $N^5|N|$ (heavier tails), with $N \sim \mathrm{N}(0,1)$. The figure shows that Φ 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. The figure also compares the right 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

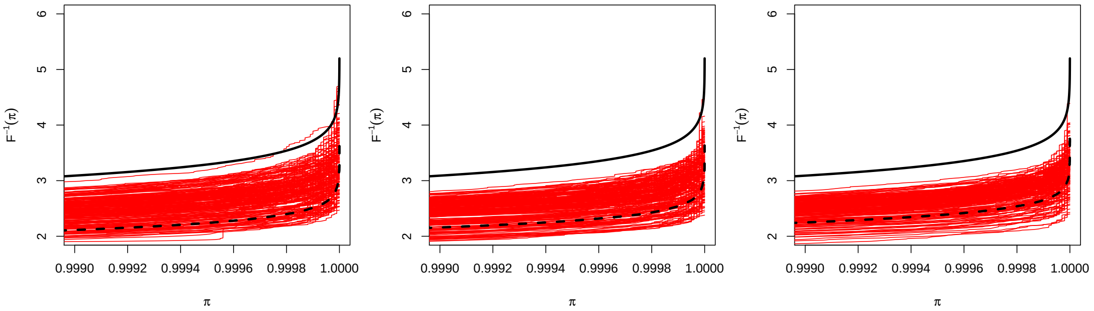

Figure 4: Inverse (F^{-1}) of the distribution function of T_0 (- - -), of the standard normal variable (—), and 200 bootstrap estimators of the distribution function of T_0 (red curves), when X is a standardised $\mathrm{F}(5,5)$, $n = 50$ (left), $n = 100$ (middle), $n = 250$ (right).

success of bootstrap in correcting this problem we compare the bootstrap and normal approximations for several skewed and heavy-tailed distributions. In particular, Figure 4 shows results obtained when $X = (U - EU)/(\mathrm{var}\, U)^{1/2}$, with $U \sim \mathrm{F}(5,5)$. 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 Figure 4 we show the right 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. The figure 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.

Note that, when implementing the bootstrap, the number B of bootstrap samples has to 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, 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_α, where $\alpha < \frac{1}{2}$, we found that one should take B no less than $100/\alpha$.

Let hc and $\mathrm{hc}_{\mathrm{norm}}$ denote, respectively, the theoretical and the normal versions of the higher criticism statistic, defined by the formula at the right hand side of (3.2), replacing there the bootstrap quantiles $\widehat{t}_\alpha^{(j)}$ by $t_\alpha^{(j)}$ and z_α, respectively, where $t_\alpha^{(j)}$

denote the $1-\alpha$ theoretical quantiles of $T_0^{(j)}$ and z_α denote the $1-\alpha$ quantile of the standard normal distribution. To illustrate the success of bootstrap in applications of the higher criticism statistic, in our simulations we compared the statistic hc which we could use if we knew the distribution F_X, the bootstrap statistic hc_n defined at (3.2), where the unknown quantiles $t_\alpha^{(j)}$ are estimated as the bootstrap quantities $\widehat{t}_\alpha^{(j)}$ as discussed in the previous paragraph, and the normal version $\mathrm{hc}_{\mathrm{norm}}$. We constructed histograms of these three versions of the higher criticism statistic, 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 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 zero, and under various alternatives $H_1^{(n)}$, where we set a fraction ϵ_n of these means equal to $\tau_n\, n^{-1/2}$, with $\tau_n > 0$. As in section 3 we took $p = n^{1/\theta}$, $\epsilon_n = n^{-\beta/\theta}$ and $\tau_n = \sqrt{2r \log p}$, where we chose β and r to be on the frontier of the $r > \rho_\theta(\beta)$.

Figure 5 shows the histograms under H_0 and under various alternatives $H_1^{(n)}$ located on the frontier ($r = \rho_\theta(\beta)$, 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 standardised $\mathrm{F}(5,5)$ variables, $n = 100$ and $\theta = \frac{1}{2}$. 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 $\mathrm{hc}_{\mathrm{norm}}$ show that the distribution of $\mathrm{hc}_{\mathrm{norm}}$ 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 rather 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 $\mathrm{hc}_{\mathrm{norm}}$.

We also compared histograms for other heavy-tailed and skewed distribution, such as the Pareto, and reached similar conclusions. Furthermore, we considered skewed but less-heavy tailed distributions, such as the chi-squared(10) distribution. There too we obtained similar results, but, while the bootstrap remained the best approximation, the normal approximation performed better than in heavy-tailed cases. We also considered values of (β, r) further away 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.

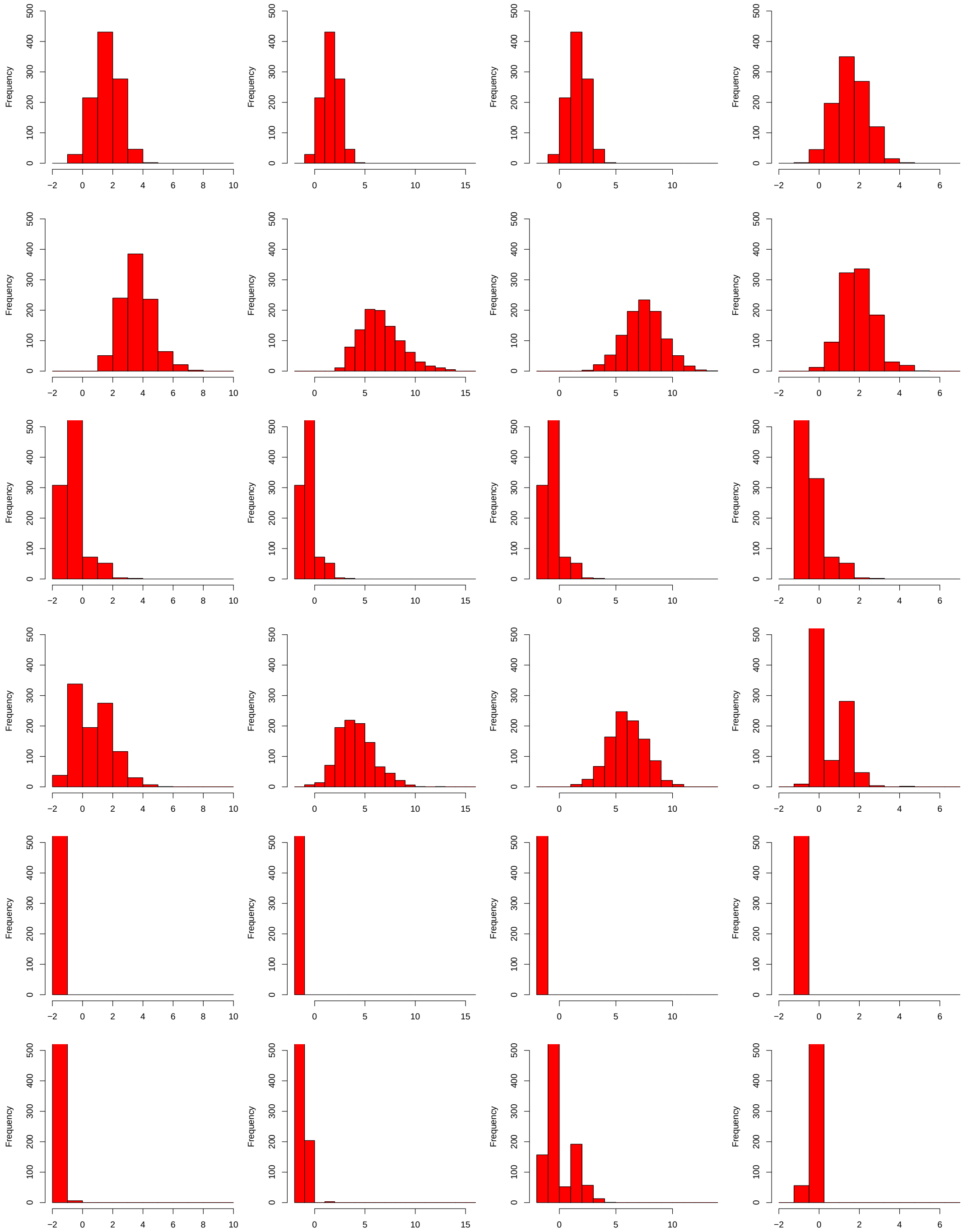

Figure 5: Histograms of hc statistics under H_0 (rows 1,3,5) or under $H_1^{(n)}$ (rows 2,4,6), when the X_j's are standardised F$(5,5)$ variables, $n = 100$, $\theta = \frac{1}{2}$, $p = n^{1/\theta}$, $\epsilon_n = n^{-\beta/\theta}$ and $\tau_n = \sqrt{2r \log p}$, where $r = \rho_\theta(\beta)$. In each row, from left to right, $\beta = \frac{1}{2}$, $\beta = \frac{1}{2} + \frac{1}{4}(1 - \theta)$, $\beta = \frac{3}{4}$ and $\beta = 1$. Rows 1 and 2 are for the theoretical hc; rows 3 and 4 for hc$_n$; and rows 5 and 6 for hc$_{\text{norm}}$.

5 Technical arguments

5.1 Preliminaries

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

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

$$\frac{P(T_c > x)}{1 - \Phi(x - c)} = \exp\big\{ - n^{-1/2} \, \tfrac{1}{6} \left(2\, x^3 - 3\, c\, x^2 + c^3\right) \gamma\big\} \\ \times \Big[1 + \theta(c, n, x) \Big\{(1 + |x|)\, n^{-1/2} + (1 + |x|)^4\, n^{-1}\Big\}\Big] \qquad (5.1)$$

as $n \to \infty$, where the function θ 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 $E|X|^4 \le B$, $E(X^2) = 1$ and $E(X) = 0$, and uniformly in c and x satisfying $0 \le x \le B\, n^{1/4}$ and $0 \le c \le u\, x$, where $0 < u < 1$.

We shall employ Theorem A to prove the theorem below. Details are given in a longer version of this paper (Delaigle *et al.*, 2010). Take $\mathcal{F}$ to be any subset of the class of distributions F of the random variable X, such that $E(|X|^{6+\epsilon}) \le B$ 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 (2.6), let $t = t_\alpha$ and $t = \widehat{t}_\alpha$ denote the respective solutions of $P(T_0 > t) = \alpha$ and $P(T_0^* > t \,|\, \mathcal{X}) = \alpha$, and recall that $z_\alpha = (1 - \Phi)^{-1}(\alpha)$. Take $\eta \in (0, \epsilon/\{4(6 + \epsilon)\})$, and let T_c and $\widehat{t}_\alpha$ denote independent random variables with the specified marginal distributions.

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

$$P_F(T_c > \widehat{t}_\alpha) = P_F(T_c > t_\alpha)\, \exp\left\{n^{-1/2}\, \tfrac{1}{6}\, c \left(3\, z_\alpha^2 - c^2\right) \gamma\right\} \\ \times \Big[1 + O\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\Big] \\ + O\Bigg[\sum_{k=1}^{3} P_F\bigg\{\bigg|\frac{1}{n} \sum_{i=1}^{n} (1 - E)\, X_i^k\bigg| > n^{-(1/4) - \eta}\bigg\} \\ + P_F\bigg\{\bigg|\frac{1}{n} \sum_{i=1}^{n} (1 - E)\, X_i^4\bigg| > B\bigg\}\Bigg] \qquad (5.2)$$

as $n \to \infty$, uniformly in all $F \in \mathcal{F}$ and in all c and z_α satisfying $0 \le z_\alpha \le B\, n^{1/4}$ and $0 \le c \le u\, z_\alpha$, where $0 < u < 1$.

5.2 Proof of Theorem 2

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

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

$$P(|Y| > y) \leq K_1 \exp\big(-K_2\, y^{\xi}\big) \tag{5.3}$$

for all $y > 0$, where $K_1, K_2, \xi > 0$, then for each $\lambda > 1$ there exist constants $K_3, K_4 > 0$, depending only on K_1, K_2, ξ and λ, such that for all $y > 0$,

$$P\bigg(\bigg|\sum_{i=1}^{n} Y_i\bigg| > y\bigg) \leq 2 \exp\bigg(-\frac{y^2}{2\,\lambda\, n}\bigg) + K_3 \exp\bigg(-\frac{y^{\xi}}{K_4}\bigg).$$

We use Theorem C to bound the remainder terms in Theorem B. If $P_F(|X| > x) \leq C_1 \exp(-C_2\, x^{\xi_1})$ and we take $Y = (1-E)\, X^k$ for an integer k, then (5.3) holds for constants K_1 and K_2 depending on C_1, C_2 and ξ_1, and with $\xi = \xi_1/k$. In particular, for all $x > 0$,

$$P_F\bigg\{\bigg|\sum_{i=1}^{n} (1-E)\, X_i^k\bigg| > x\,\big(\mathrm{var}\, X^k\big)^{1/2}\bigg\} \leq 2 \exp\bigg(-\frac{x^2}{2\,\lambda\, n}\bigg) + K_3 \exp\bigg(-\frac{x^{\xi_1/k}}{K_4}\bigg).$$

Taking $k = 1$, 2 or 3, and $x = x_{kn} = \mathrm{const.}\, n^{(3/4)-\eta_1}$ for some $\eta_1 > 0$; or $k = 4$ and $x = x_{kn} = \mathrm{const.}$; we deduce that in each of these settings,

$$P_F\bigg\{\bigg|\frac{1}{n}\sum_{i=1}^{n} (1-E)\, X_i^k\bigg| > x_{nk}\bigg\} = \begin{cases} O\big\{\exp\big(-n^{(3\xi_1/4k)-\eta_2}\big)\big\} & \text{if } k = 1, 2, 3 \\ O\big\{\exp\big(-n^{\xi_1/4}/K_5\big)\big\} & \text{if } k = 4\,, \end{cases}$$

where $\eta_2 > 0$ decreases to zero as $\eta_1 \downarrow 0$. Therefore the $O[\ldots]$ remainder term in (5.2) equals $O\{\exp(-n^{(3\xi_1/16)-\eta_2})\}$, and so Theorem 2 is implied by Theorem B.

5.3 Proof of (2.12)

Note that, by Theorem B in section 5.1, Theorems 1 and 2 continue to hold if we replace the left-hand sides of (2.8) and (2.10) by $P_F(T_c > \widehat{t}_\alpha)$, provided we also replace the factor α on the right-hand sides by $P_F(T_c > t_\alpha)$. The uniformity with which (2.8) and (2.10) hold now extends (in view of Theorem B) to c such that $0 \leq c \leq u\, z_\alpha$ with $0 < u < 1$, as well as to α satisfying $0 \leq z_\alpha \leq B\, n^{1/4}$.

Acknowledgement

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

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. Roy. Statist. Soc.* Ser. B **57**, 289–300.

BERNHARD, G., KLEIN, M. AND HOMMEL, G. (2004). Global and multiple test procedures using ordered p-values — a review. *Statist. Papers* **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., STRODE, 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.* **102**, 4080–4084.

DELAIGLE, A. AND HALL, P. (2009). Higher criticism in the context of unknown distribution, non-independence and classification. In *Perspectives in Math-*

ematical Sciences I: Probability and Statistics, 109–138. Eds N. Sastry, M. Delampady, B. Rajeev and T.S.S.R.K. Rao. World Scientific.

DELAIGLE, A., HALL, P. AND JIN, J. (2010). Robustness and accuracy of methods for high dimensional data analysis based on Student's t statistic–long version. Available at `http://arxiv.org/`

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.

DUDOIT, 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. Amer. Statist. Assoc.* **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. Amer. Statist. Assoc.* **102**, 1282–1288.

FAN, J. AND LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc.* Ser. 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. Multivariate 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*, to appear.

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.

JIN, J. (2007). Proportion of nonzero normal means: universal oracle equivalences and uniformly consistent estimators. *J. Roy. Statist. Soc.* Ser. 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. Amer. Statist. Assoc.* **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. *Brit. J. Math. Statist. Psych.* **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.* American College of Physicians, Philadelphia.

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. Verojatnost. i Primenen* **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. `http://arxiv.org/pdf/q-bio/0402043v1`

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

PIGEOT, I. (2000). Basic concepts of multiple tests — A survey. *Statist. Papers* **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. Theoret. Probab.* **12**, 385–398.

STOREY, J.D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc.* Ser. 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. *Systems and Computers in Japan* **32**, 16–24.

TAMHANE, A.C. AND DUNNETT, C.W. (1999). Stepwise multiple test procedures with biometric applications. *J. Statist. Plann. Inf.* **82**, 55–68.

WANG, Q. AND HALL, P. (2009). Relative errors in central limit theorems for Student's t statistic, with applications. *Statist. Sinica* **19**, 343–354.

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

A PAGES 29–39: NOT-FOR-PUBLICATION APPENDIX

A.1 Proof of Theorem B

Step 1: Expansions of t_α and $\widehat{t}_\alpha$. The main results here are (A.3) and (A.5). To derive them, take W^* to have the distribution of $(X_i^* - \bar{X})/S$, where S is as in (2.1), and, for $k = 3$ and 4, put

$$\widehat{\gamma}_k = E\big(W^{*k} \,|\, \mathcal{X}\big) = \frac{1}{n\, S^k} \sum_{i=1}^{n} (X_i - \bar{X})^k\,,$$

where $W^{*k} = (W^*)^k$. Letting $c = 0$ in Theorem A, and taking X there to have the distribution of W^* conditional on $\mathcal{X}$, we deduce that if $B > 1$ is given,

$$\frac{P_F(T_0^* > x \,|\, \mathcal{X})}{1 - \Phi(x)} = \exp\big(- n^{-1/2}\, \tfrac{1}{3}\, x^3\, \widehat{\gamma}\big) \times \Big[1 + \Theta_1(n,x) \Big\{(1+|x|)\, n^{-1/2} + (1+|x|)^4\, n^{-1}\Big\}\Big]\,, \qquad \text{(A.1)}$$

where $\widehat{\gamma} = \widehat{\gamma}_3$ and:

> the random function $\Theta_1(n,x)$ satisfies $|\Theta_1(n,x)| \le C_1(B)$ (where $C_1(B)$ is the same constant introduced in Theorem A) uniformly in datasets $\mathcal{X}$ for which $S > \frac{1}{2}$ and $\widehat{\gamma}_4 \le B$, and uniformly also in x satisfying $0 \le x \le B\, n^{1/4}$. (A.2)

Properties (A.1) and (A.2) imply that$\widehat{t}_\alpha$ satisfies:

$$\widehat{t}_\alpha = z_\alpha \Big[1 - \tfrac{1}{3}\, \widehat{\gamma}\, n^{-1/2}\, z_\alpha + \Theta_2(n,\alpha) \big\{(1+z_\alpha)^{-1}\, n^{-1/2} + (1+z_\alpha)^2\, n^{-1}\big\}\Big]\,, \qquad \text{(A.3)}$$

where $z = z_\alpha$ is the solution of $1 - \Phi(z_\alpha) = \alpha$ and, in the case $j = 2$:

> the random function $\Theta_j(n,\alpha)$ satisfies $|\Theta_j(n,\alpha)| \le C_j(B)$ (where $C_j(B)$ is a finite, positive constant) uniformly in datasets $\mathcal{X}$ for which $S > \frac{1}{2}$ and $\widehat{\gamma}_4 \le B$, and uniformly also in α satisfying $\frac{1}{2} \le 1-\alpha \le 1-\Phi(B\, n^{1/4})$ (A.4)

Analogously, Theorem A implies that t_α satisfies:

$$t_\alpha = z_\alpha \Big[1 - \tfrac{1}{3}\, \gamma\, n^{-1/2}\, z_\alpha + \theta(n,\alpha) \big\{(1+z_\alpha)^{-1}\, n^{-1/2} + (1+z_\alpha)^2\, n^{-1}\big\}\Big]\,, \qquad \text{(A.5)}$$

where $z = z_\alpha$ is the solution of $1 - \Phi(z_\alpha) = \alpha$ and:

the function $\theta(n,\alpha)$ satisfies $|\theta(n,\alpha)| \le C_2(B)$ (with $C_2(B)$ denoting a finite, positive constant) uniformly in distributions of X for which $E(X)=0$, $E(X^2)=1$ and $E(X^4) \le B$, and uniformly also in α satisfying $0 \le z_\alpha \le B\, n^{1/4}$. (A.6)

The derivations of the pairs of properties (A.3) and (A.4), and (A.5) and (A.6), are similar. For example, suppose that if $\widehat{t}_\alpha$ is given by (A.3) rather than by $P(T_0^* > \widehat{t}_\alpha \,|\, \mathcal{X}) = \alpha$, and that the function Θ_2 in (A.3) is open to choice except that it should satisfy (A.4). If we define $\rho(z) = z\,\{1-\Phi(z)\}/\phi(z) = 1 - z^{-2} + 3\,z^{-4} - \ldots$, then by (A.1), (A.3) and (A.4),

$$\begin{aligned} P_F(T_0^* > \widehat{t}_\alpha \,|\, \mathcal{X}) &= \{1-\Phi(\widehat{t}_\alpha)\}\, \exp\big(-n^{-1/2}\,\tfrac13\,\widehat{t}_\alpha^{\,3}\,\widehat{\gamma}\big) \\ &\qquad \times \Big[1 + \Theta_1(n,\widehat{t}_\alpha)\,\Big\{(1+|\widehat{t}_\alpha|)\,n^{-1/2} + (1+|\widehat{t}_\alpha|)^4\,n^{-1}\Big\}\Big] \\ &= (2\pi)^{-1/2}\,\exp\big\{-\tfrac12\,z_\alpha^2\,\big(1-\tfrac23\,\widehat{\gamma}\,n^{-1/2}\,z_\alpha\big) - n^{-1/2}\,\tfrac13\,z_\alpha^3\,\widehat{\gamma}\big\} \\ &\qquad \times z_\alpha^{-1}\,\rho(z_\alpha)\,\Big[1+\Theta_3(n,z_\alpha)\,\Big\{(1+z_\alpha)\,n^{-1/2} + (1+z_\alpha)^4\,n^{-1}\Big\}\Big] \\ &= \{1-\Phi(z_\alpha)\}\,\Big[1+\Theta_3(n,z_\alpha)\,\Big\{(1+z_\alpha)\,n^{-1/2} + (1+z_\alpha)^4\,n^{-1}\Big\}\Big]\,, \end{aligned} \tag{A.7}$$

where Θ_3 satisfies (A.4). By judicious choice of Θ_2, satisfying (A.4), we can ensure that Θ_3 in (A.7) vanishes, up to the level of discreteness of the conditional distribution function of T_0^*. In this case the right-hand side of (A.7) equals simply $1-\Phi(z_\alpha)=\alpha$, so that $\widehat{t}_\alpha$ indeed has the intended property, i.e. $P(T_0^* > \widehat{t}_\alpha \,|\, \mathcal{X}) = \alpha$.

Step 2: Expansions of the difference between $\widehat{t}_\alpha$ and t_α. The main results here are (A.10) and (A.11). To obtain them, first combine (A.3) and (A.5) to deduce that:

$$\widehat{t}_\alpha - t_\alpha = \tfrac13\,z_\alpha^2\,(\gamma-\widehat{\gamma})\,n^{-1/2} + \Theta_4(n,\alpha)\,\big\{n^{-1/2} + (1+z_\alpha)^3\,n^{-1}\big\}\,, \tag{A.8}$$

where, for $j=4$:

the random function $\Theta_j(n,\alpha)$ satisfies $|\Theta_j(n,\alpha)| \le C_j(B)$ (with $C_j(B)$ denoting a finite, positive constant) uniformly in datasets $\mathcal{X}$ for which $S > \frac12$ and $\widehat{\gamma}_4 \le B$; uniformly in distributions of X for which $E(X)=0$, $E(X^2)=1$ and $E(X^4) \le B$; and uniformly also in α satisfying $0 \le z_\alpha \le B\,n^{1/4}$. (A.9)

Using (A.5), (A.6), (A.8) and (A.9) we deduce that:

$$\begin{aligned} t_\alpha^2\,(\widehat{t}_\alpha - t_\alpha) &= \tfrac13\,z_\alpha^4\,n^{-1/2}\,(\gamma-\widehat{\gamma}) + \Theta_5(n,\alpha)\,\big\{(1+z_\alpha)^2\,n^{-1/2} + (1+z_\alpha)^5\,n^{-1}\big\}\,, \\ t_\alpha\,(\widehat{t}_\alpha - t_\alpha)^2 &= \tfrac19\,z_\alpha^5\,n^{-1}\,(\gamma-\widehat{\gamma})^2 + \Theta_6(n,\alpha)\,\big\{(1+z_\alpha)^3\,n^{-1} + (1+z_\alpha)^6\,n^{-3/2}\big\} \end{aligned}$$

and $(\widehat{t}_\alpha - t_\alpha)^3 = \Theta_7(n, \alpha)\,(1 + z_\alpha)^6\, n^{-3/2} \le \Theta_8(n, \alpha)\,(1 + z_\alpha)^4\, n^{-1}$, where $\Theta_5, \ldots, \Theta_9$ (the latter appearing below) satisfy (A.9). Therefore,

$$\begin{aligned} \widehat{t}_\alpha^3 - t_\alpha^3 &= 3\,t_\alpha^2\,(\widehat{t}_\alpha - t_\alpha) + 3\,t_\alpha\,(\widehat{t}_\alpha - t_\alpha)^2 + (\widehat{t}_\alpha - t_\alpha)^3 \\ &= z_\alpha^4\, n^{-1/2}\,(\gamma - \widehat{\gamma}) + \Theta_9(n, \alpha)\,\big\{(1 + z_\alpha)^2\, n^{-1/2} + (1 + z_\alpha)^5\, n^{-1}\big\}. \end{aligned} \tag{A.10}$$

Similarly, using (A.5) and (A.8),

$$\begin{aligned} (\widehat{t}_\alpha - c)^2 - (t_\alpha - c)^2 &= 2\,(t_\alpha - c)\,(\widehat{t}_\alpha - t_\alpha) + (\widehat{t}_\alpha - t_\alpha)^2 \\ &= \tfrac{2}{3}\,(z_\alpha - c)\, z_\alpha^2\,(\gamma - \widehat{\gamma})\, n^{-1/2} \\ &\qquad + \Theta_{10}(c, n, \alpha)\,\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\,, \end{aligned} \tag{A.11}$$

where, for $j \ge 10$:

> the random function $\Theta_j(c, n, \alpha)$ satisfies $|\Theta_j(c, n, \alpha)| \le C_j(B)$ (with $C_j(B)$ denoting a finite, positive constant) uniformly in datasets $\mathcal{X}$ for which $S > \frac{1}{2}$ and $\widehat{\gamma}_4 \le B$, uniformly in distributions of X for which $E(X) = 0$, $E(X^2) = 1$ and $E(X^4) \le B$, and uniformly also in c such that $0 \le c \le u\, z_\alpha$ where $0 < u < 1$, and in α such that $0 \le z_\alpha \le B\, n^{1/4}$. (A.12)

Step 3: Initial expansion of $P(T_c > \widehat{t}_\alpha)$. To derive (A.16), the main result in this step, note that by (A.3)–(A.5) and (A.11),

$$\begin{aligned} 1 - \Phi(\widehat{t}_\alpha - c) &= (z_\alpha - c)^{-1}\,\rho(z_\alpha - c)\,(2\pi)^{-1/2}\,\exp\big\{-\tfrac{1}{2}\,(\widehat{t}_\alpha - c)^2\big\} \\ &\qquad \times \Big[1 + \Theta_{11}(c, n, \alpha)\,\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^2\, n^{-1}\big\}\Big] \\ &= (z_\alpha - c)^{-1}\,\rho(z_\alpha - c)\,(2\pi)^{-1/2} \\ &\qquad \times \exp\Big\{-\tfrac{1}{2}\,(t_\alpha - c)^2 - \tfrac{1}{3}\,(z_\alpha - c)\, z_\alpha^2\,(\gamma - \widehat{\gamma})\, n^{-1/2}\Big\} \\ &\qquad \times \Big[1 + \Theta_{12}(c, n, \alpha)\,\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\Big] \\ &= \{1 - \Phi(t_\alpha - c)\}\,\exp\big\{-\tfrac{1}{3}\,(z_\alpha - c)\, z_\alpha^2\,(\gamma - \widehat{\gamma})\, n^{-1/2}\big\} \\ &\qquad \times \Big[1 + \Theta_{13}(c, n, \alpha)\,\big\{(1 + z_\alpha)\, n^{-1/2} + (1 + z_\alpha)^4\, n^{-1}\big\}\Big]\,. \end{aligned} \tag{A.13}$$

If T_c is statistically independent of $\widehat{t}_\alpha$ then, by (5.1), (A.10) and (A.11) (the latter with $c = 0$),

$$\frac{P_F(T_c > \widehat{t}_\alpha \,|\, \widehat{t}_\alpha)}{1 - \Phi(\widehat{t}_\alpha - c)}$$

$$
\begin{aligned}
&= \exp\big\{ - n^{-1/2}\, \tfrac{1}{6}\, \big(2\, \widehat{t}_\alpha^{\,3} - 3\, c\, \widehat{t}_\alpha^{\,2} + c^3\big)\, \gamma \big\} \\
&\qquad \times \Big[1 + \Theta_{14}(c,n,\alpha) \Big\{ (1+z_\alpha)\, n^{-1/2} + (1+z_\alpha)^4\, n^{-1} \Big\} \Big] \\
&= \exp\big\{ - n^{-1/2}\, \tfrac{1}{6}\, \big(2\, t_\alpha^3 - 3\, c\, t_\alpha^2 + c^3\big)\, \gamma \big\} \\
&\qquad \times \exp\Big[- n^{-1}\, \tfrac{1}{3}\, \gamma\, (\gamma - \widehat{\gamma})\, \big\{ z_\alpha^4 - c\, z_\alpha^3 \big\} \Big] \\
&\qquad \times \Big[1 + \Theta_{15}(c,n,\alpha) \Big\{ (1+z_\alpha)\, n^{-1/2} + (1+z_\alpha)^4\, n^{-1} \Big\} \Big] \\
&= \frac{P_F(T_c > t_\alpha)}{1 - \Phi(t_\alpha - c)} \Big[1 + \Theta_{16}(c,n,\alpha) \Big\{ (1+z_\alpha)\, n^{-1/2} + (1+z_\alpha)^4\, n^{-1} \Big\} \Big] . \qquad \text{(A.14)}
\end{aligned}
$$

Combining (A.13) and (A.14) we deduce that:

$$
\begin{aligned}
P_F(T_c > \widehat{t}_\alpha \,|\, \widehat{t}_\alpha) &= P_F(T_c > t_\alpha) \cdot \exp\big\{ - \tfrac{1}{3}\, (z_\alpha - c)\, z_\alpha^2\, (\gamma - \widehat{\gamma})\, n^{-1/2} \big\} \\
&\qquad \times \Big[1 + \Theta_{17}(c,n,\alpha) \big\{ (1+z_\alpha)\, n^{-1/2} + (1+z_\alpha)^4\, n^{-1} \big\} \Big] . \qquad \text{(A.15)}
\end{aligned}
$$

Reflecting (A.12), let $\mathcal{G}_1(B)$ denote the class of distribution functions F of X such that $E(X) = 0$, $E(X^2) = 1$ and $E(X^4) \le B$; write P_F for probability measure when $\mathcal{X}$ is drawn from the population with distribution function $F \in \mathcal{G}_1$; let $\mathcal{D}$ denote any given event, shortly to be defined concisely; let $\mathcal{E}(B)$ be the intersection of $\mathcal{D}$ and the events $S > \frac{1}{2}$ and $\widehat{\gamma}_4 \le B$; and write $\widetilde{\mathcal{E}}(B)$ for the complement of $\mathcal{E}(B)$. In view of (A.15),

$$
\begin{aligned}
P_F(T_c > \widehat{t}_\alpha) &= P_F(T_c > t_\alpha) \cdot E\Big[\exp\Big\{ n^{-1/2}\, \tfrac{1}{3}\, (z_\alpha - c)\, z_\alpha^2\, (\widehat{\gamma} - \gamma) \Big\}\, I\{\mathcal{E}(B)\} \Big] \\
&\qquad \times \Big[1 + O\big\{ (1+z_\alpha)\, n^{-1/2} + (1+z_\alpha)^4\, n^{-1} \big\} \Big] \\
&\qquad + O\big[P_F\big\{ \widetilde{\mathcal{E}}(B) \big\} \big] , \qquad \text{(A.16)}
\end{aligned}
$$

uniformly in the following sense:

uniformly in $F \in \mathcal{G}_1(B)$, in c such that $0 \le c \le u\, z_\alpha$, where $0 < u < 1$, and in α such that $0 \le z_\alpha \le B\, n^{1/4}$. (A.17)

Step 4: Simplification of right-hand side of (A.16). Here we derive a simple formula, (A.28), for the expectation on the right-hand side of (A.16). That result, when combined with (A.16) and (A.17), leads quickly to Theorem B.

Put $\Delta_k = n^{-1} \sum_i (X_i^k - EX^k)$, write $\mathcal{D}_k$ to denote the event that $|\Delta_k| \le C_3\, n^{-(1/4)-\eta}$ where $C_3 > 0$ and $\eta \in (0, \frac{1}{2})$, and put $\mathcal{D} = \mathcal{D}_1 \cap \mathcal{D}_2 \cap \mathcal{D}_3$. Observe that

$$
\widehat{\gamma} = \big(\gamma + \Delta_3 - 3\, \Delta_1\, \Delta_2 - 3\Delta_1 + 2\, \Delta_1^3 \big) \big/ \big(1 + \Delta_2 - \Delta_1^2 \big) , \qquad \text{(A.18)}
$$

From this property it can be proved that if $|\gamma| \le B$ and C_3 is sufficiently small, depending only on B, then $|\widehat{\gamma} - \gamma| \le n^{-(1/4)-\eta}$ whenever $\mathcal{D}$ holds. Therefore if $\mathcal{D}$ holds, and $0 \le c \le u\, z_\alpha$ for $0 < u < 1$, and $0 \le z_\alpha \le B\, n^{1/4}$, then

$$n^{-1/2}\, |z_\alpha - c|\, z_\alpha^2\, |\widehat{\gamma} - \gamma| \le B^3\, n^{-\eta} \,.$$

In these circumstances, defining $\Delta = n^{-1/2}\, \frac{1}{3}\, (z_\alpha - c)\, z_\alpha^2\, (\widehat{\gamma} - \gamma)$, we have:

$$\left| e^{\Delta} - \sum_{j=0}^{r} \frac{\Delta^j}{j!} \right| I(\mathcal{D}) \le \big(B^3\, n^{-\eta}\big)^{r+1} \exp\big(B^3\, n^{-\eta}\big) \,. \tag{A.19}$$

Note too that if $E(X^6) \le B$ then

$E(\Delta_{k_1}^{r_1}\, \Delta_{k_2}^{r_2}) \le C_4(B)\, n^{-1}$ whenever k_1 and k_2 take values in the set $\{1, 2, 3\}$, r_1 and r_2 are nonnegative, and $r_1 + r_2 = 1$ or 2. (A.20)

Also, in the same context as (A.20), if $r_1 + r_2 = 2$ then

$$E\big\{|\Delta_{k_1}^{r_1}\, \Delta_{k_2}^{r_2}|\, I\big(\widetilde{\mathcal{D}}\big)\big\} \le E\big(|\Delta_{k_1}^{r_1}\, \Delta_{k_2}^{r_2}|\big) \le \big\{E\big(\Delta_{k_1}^{2r_1}\big)\, E\big(\Delta_{k_2}^{2r_1}\big)\big\}^{1/2} \le C_4(B)\, n^{-1} \,; \tag{A.21}$$

and if $r_1 = 1$ and $r_2 = 0$, and η is sufficiently small,

$$\begin{aligned} E\big\{|\Delta_{k_1}^{r_1}\, \Delta_{k_2}^{r_2}|\, I\big(\widetilde{\mathcal{D}}\big)\big\} \le \big\{E\Delta_{k_1}^{2r_1}\, P\big(\widetilde{\mathcal{D}}\big)\big\}^{1/2} &\le C_5(B, \eta)\, \big(n^{-1}\, n^{-(1/2)-\zeta}\big)^{1/2} \\ &= C_5(B, \eta)\, n^{-(3/4)-(\zeta/2)} \end{aligned} \tag{A.22}$$

where $\zeta = \zeta(\eta) > 0$. In deriving (A.22) we used the fact that $P(\widetilde{\mathcal{D}}) \le P(\widetilde{\mathcal{D}}_1) + P(\widetilde{\mathcal{D}}_2) + P(\widetilde{\mathcal{D}}_3)$, and that, by Markov's inequality (employing the fact that $E|X|^{6+\epsilon} < \infty$ and choosing $\eta < \epsilon/\{4(3 + \epsilon)\}$),

$$\begin{aligned} P\big(\widetilde{\mathcal{D}}_k\big) &\le \big(C_3\, n^{-(1/4)-\eta}\big)^{-\{2+(\epsilon/3)\}}\, E\big(|\Delta_k|^{2+(\epsilon/3)}\big) \\ &\le C_6(B, \eta)\, n^{(1/2)+2\eta+(\epsilon/12)+(\eta\epsilon/3)-\{2+(\epsilon/3)\}/2} \le C_6(B, \eta)\, n^{-(1/2)-\zeta} \end{aligned}$$

for $k = 1, 2, 3$, where $\zeta > 0$. Therefore,

$$P\big(\widetilde{\mathcal{D}}\big) \le 3\, C_6(B, \eta)\, n^{-(1/2)-\zeta} \,. \tag{A.23}$$

If $r_1 + r_2 + r_3 \ge 3$ then an argument similar to that leading to (A.23) shows that

$$E\big\{\big|\Delta_1^{r_1}\, \Delta_2^{r_2}\, \Delta_3^{r_3}\big|\, I(\mathcal{D})\big\} \le C_7(B, \eta)\, \big(n^{-1/2}\big)^2\, \big(n^{-(1/4)-\eta}\big)^{r_1+r_2+r_3-2} \,. \tag{A.24}$$

Combining (A.20), (A.21), (A.22) and (A.24); using Taylor expansion to derive approximations to $\widehat{\gamma} - \gamma$, starting from (A.18); noting the definition of Δ given in the previous paragraph; and observing that $n^{-1/2}\,|z_\alpha - c|\,z_\alpha^2 \le B^3\,n^{1/4}$ if $0 \le c \le u\,z_\alpha$ with $0 < u < 1$, and $0 \le z_\alpha \le B\,n^{1/4}$; we deduce that:

$$\big|E\{\Delta^j\, I(\mathcal{D})\}\big| \le \begin{cases} C_8(B,j)\,n^{1/4}\,n^{-1} & \text{if } j=1 \\ C_8(B,j)\,\big(n^{1/4}\big)^j\,n^{-1}\,\big(n^{-(1/4)-\eta}\big)^{j-2} & \text{if } j \ge 2 \end{cases}$$
$$\le C_8(B,j)\,n^{-1/2}\,. \tag{A.25}$$

Using (A.19), (A.23) and (A.25), and choosing r to be the least integer such that $(r+1)\,\eta \ge \frac{1}{2}$, we deduce that:

$$E\Big[\exp\Big\{n^{-1/2}\,\tfrac{1}{3}\,(z_\alpha - c)\,z_\alpha^2\,(\widehat{\gamma}-\gamma)\Big\}\,I(\mathcal{D})\Big] = 1 + O\big(n^{-1/2}\big)\,, \tag{A.26}$$

uniformly in the following sense:

uniformly in $F \in \mathcal{G}_2(B)$, in c such that $0 \le c \le u\,z_\alpha$ with $0 < u < 1$, and in α such that $0 \le z_\alpha \le B\,n^{1/4}$, (A.27)

where $\mathcal{G}_2(B)$ denotes the intersection of $\mathcal{G}_1(B)$ (defined at (A.17)) with the class of distributions of X such that $E(|X|^{6+\epsilon}) \le B$.

An argument almost identical to that leading to (A.26) and (A.27) shows that the same pair of results holds if we replace $\mathcal{D}$ by the event $\mathcal{D}_1$ that $S > \frac{1}{2}$ and $\widehat{\gamma}_4 \le B$. The only change needed is the observation that, since $F \in \mathcal{G}_1(B)$ entails $E(|X|^{6+\epsilon}) \le B$, $P(\widetilde{\mathcal{D}}_1)$ is uniformly bounded above by a constant multiple of $n^{-1/2}$. This follows from the fact that, if $Y_1, Y_2, \ldots$ are random variables satisfying $E|Y|^{6+\epsilon} < \infty$, then $P\{|\sum_{i\le n}(1-E)\,Y_i^4| > n\} \le$ const. $n^{-(1/2)-(\epsilon/4)}$. Therefore, in the argument in (A.22) we can replace the bound const. $n^{-(1/2)-\zeta}$ to $P(\widetilde{\mathcal{D}})$ by the bound const. $n^{-(1/2)-(\epsilon/4)}$ to $P(\widetilde{\mathcal{D}}_1)$. This means that (A.26) holds if we replace $\mathcal{D}$ there by the event $\mathcal{D} \cap \mathcal{D}_1$, i.e. the event $\mathcal{E}(B)$ introduced just above (A.16). That is,

$$E\Big[\exp\Big\{n^{-1/2}\,\tfrac{1}{3}\,(z_\alpha - c)\,z_\alpha^2\,(\widehat{\gamma}-\gamma)\Big\}\,I\{\mathcal{E}(B)\}\Big] = 1 + O\big(n^{-1/2}\big)\,, \tag{A.28}$$

uniformly in the sense of (A.27).

Together, (A.16), (A.17) and (A.28) imply that (5.2) holds uniformly in $F \in \mathcal{F}$, in c such that $0 \le c \le u\,z_\alpha$, with $0 < u < 1$ and in α such that $0 \le z_\alpha \le B\,n^{1/4}$, completing the proof of Theorem B.

A.2 Proof of Theorem 3

Throughout this proof we use the notation $\mathrm{hc}_n^* = \mathrm{hc}_n(\alpha_0^*)$, where $\alpha_0^* = n\,p^{-1}\log p$ denotes the value of α_0 stated in the theorem. Also, for two positive sequences a_n and b_n, we write $a_n \lesssim b_n$ when $\limsup_{n\to\infty}(a_n/b_n) \le 1$. We use the equivalent notation $b_n \gtrsim a_n$.

Fix $\alpha \in (0,1)$. Let $G_p(\alpha) = p^{-1}\sum_{j=1}^p I\big(T_{c^{(j)}}^{(j)} > \widehat{t}_\alpha^{(j)}\big)$, and $\mathrm{hc}_{n,\alpha} = \sqrt{p}\{\alpha(1-\alpha)\}^{-1/2}\{G_p(\alpha) - \alpha\}$. We have $\mathrm{hc}_n^* = \max_{\alpha=i/p, 1\le i\le \alpha_0^* p}\mathrm{hc}_{n,\alpha}$. We introduce a non-stochastic counterpart

$$\widetilde{\mathrm{hc}}_n^* = \max_{\alpha=i/p,1\le i\le\alpha_0^*p}\widetilde{\mathrm{hc}}_{n,\alpha}$$

of hc_n^*, where $\widetilde{\mathrm{hc}}_{n,\alpha} = \sqrt{p}\{\alpha(1-\alpha)\}^{-1/2}\{\bar{G}_p(\alpha) - \alpha\}$ and $\bar{G}_p(\alpha) = p^{-1}\sum_{j=1}^p P\big(T_{c^{(j)}}^{(j)} > \widehat{t}_\alpha^{(j)}\big)$. Note that $\bar{G}_p(\alpha) = E\{G_p(\alpha)\}$.

The keys for the proofs are:

(A) There is a constant $C > 0$ such that

$$\lim_{n\to\infty} P\Big\{|\mathrm{hc}_n^* - \widetilde{\mathrm{hc}}_n^*| \le C\log p\Big\} = 1, \qquad \text{under } H_0,$$
$$\lim_{n\to\infty} P\Big\{|\mathrm{hc}_n^* - \widetilde{\mathrm{hc}}_n^*| \le C\log p\sqrt{1+\widetilde{\mathrm{hc}}_n^*}\Big\} = 1, \qquad \text{under } H_1^{(n)}.$$

(B) Under H_0, there is a constant $C > 0$ such that $\widetilde{\mathrm{hc}}_n^* \le C\log p$ for sufficiently large n.

(C) Under $H_1^{(n)}$, $\widetilde{\mathrm{hc}}_n^* = L_p p^{\delta(\beta,r,\theta)}$.

Combining **(A)**–**(B)**, there exit constants $C_1 > 0$ and $C_2 > 0$ such that $\widetilde{\mathrm{hc}}_n^* \le C_1\log p$ and $P\{|\mathrm{hc}_n^* - \widetilde{\mathrm{hc}}_n^*| \le C_2\log p\} = 1 + o(1)$. Therefore,

$$P\Big\{\mathrm{hc}_n^* \le (C_1+C_2)\log p\Big\} \ge P\Big\{|\mathrm{hc}_n^* - \widetilde{\mathrm{hc}}_n^*| \le C_2\log p\Big\} = 1 + o(1),$$

and part (a) of Theorem 3 follows. Combining **(A)** and **(C)** gives that

$$P\Big\{\mathrm{hc}_n^* \ge L_p p^{\delta(\beta,r,\theta)}\Big\} \ge P\Big\{\mathrm{hc}_n^* \ge \widetilde{\mathrm{hc}}_n^* - C\log p\sqrt{1+\widetilde{\mathrm{hc}}_n^*}\Big\} \to 1 \text{ as } n\to\infty,$$

and part (b) of the theorem follows. Note that C and L_p may stand for different quantities in different occurrence.

We now show **(A)**–**(C)**. Below, whenever we refer to α, we assume that $p^{-1} \le \alpha \le \alpha_0^*$. By definition, $\bar{G}_p(\alpha) = p^{-1}\sum_{j=1}^p P(T_{c^{(j)}}^{(j)} > \widehat{t}_\alpha^{(j)})$, where the fraction of $c^{(j)} = 0$

is 1 under the null and $(1-\epsilon_n)$ under the alternative. Using Theorem 1 and noting that $O(n^{-D_1}) = o(1/p)$ and that $z_\alpha \leq O(\sqrt{\log p})$ in (2.8), we have

$$P(T_{c^{(j)}}^{(j)} > \widehat{t}_\alpha^{(j)}) = \alpha\{1 + O(\sqrt{\log p}/\sqrt{n})\} + o(1/p), \qquad \text{when } c^{(j)} = 0. \tag{A.29}$$

It follows that both under the null and under the alternative,

$$(1-\epsilon_n)\alpha \lesssim \bar{G}_p(\alpha) \lesssim (1-\epsilon_n)\alpha + \epsilon_n. \tag{A.30}$$

As a result, uniformly in $\alpha \in [1/p, \alpha_0^*]$,

$$\alpha = o(1), \qquad \bar{G}_p(\alpha) = o(1), \qquad p\,\bar{G}_p(\alpha) \gtrsim p\,\alpha \geq 1. \tag{A.31}$$

Consider **(A)**. Note that for any integer $N \geq 1$ and any positive sequences a_i and b_i, $\max_{1\leq i\leq N}\{a_i b_i\} \leq \max_{1\leq i\leq N}\{a_i\}\cdot\max_{1\leq i\leq N}\{b_i\}$. By the definition of hc_n^* and $\widetilde{\mathrm{hc}}_n^*$,

$$|\mathrm{hc}_n^* - \widetilde{\mathrm{hc}}_n^*| \leq \max_{\alpha = i/p: 1\leq i\leq \alpha_0^* p} \frac{\sqrt{p}\ |G_p(\alpha) - \bar{G}_p(\alpha)|}{\sqrt{\alpha(1-\alpha)}} \leq I\cdot II,$$

where I is stochastic and II is deterministic, and

$$I = \max_{\alpha = i/p: 1\leq i\leq \alpha_0^* p} \frac{\sqrt{p}\,|G_p(\alpha) - \bar{G}_p(\alpha)|}{\sqrt{\bar{G}_p(\alpha)(1-\bar{G}_p(\alpha))}}, \qquad II = \max_{\alpha = i/p: 1\leq i\leq \alpha_0^* p} \frac{|\bar{G}_p(\alpha)(1-\bar{G}_p(\alpha))|}{\sqrt{\alpha(1-\alpha)}}.$$

To show **(A)**, it is sufficient to show that both under the null and the alternative,

$$P(I \geq C\log p) = o(1), \tag{A.32}$$

and that

$$II \lesssim 1 \quad \text{under } H_0, \qquad II \lesssim \sqrt{1 + |\widetilde{\mathrm{hc}}_n^*|} \quad \text{under } H_1^{(n)}. \tag{A.33}$$

Consider (A.32). Note that

$$P(I > C\log p) \leq \sum_{\alpha = i/p, 1\leq i\leq \alpha_0^* p} P\left\{\frac{\sqrt{p}\,|G_p(\alpha) - \bar{G}_p(\alpha)|}{\sqrt{\bar{G}_p(\alpha)(1-\bar{G}_p(\alpha))}} > C\log(p)\right\}. \tag{A.34}$$

For each α, applying Bennett's inequality [Shorack and Wellner (1986) page 851] with $X_j = I(T_{c^{(j)}}^{(j)} > \widehat{t}_\alpha^{(j)}) - P(T_{c^{(j)}}^{(j)} > \widehat{t}_\alpha^{(j)})$ and $\lambda = C\log(p)\sqrt{\bar{G}_p(\alpha)(1-\bar{G}_p(\alpha))}$,

$$\begin{aligned} P\left\{\frac{\sqrt{p}\,|G_p(\alpha) - \bar{G}_p(\alpha)|}{\sqrt{\bar{G}_p(\alpha)\{1-\bar{G}_p(\alpha)\}}} C\log p\right\} &\equiv P\{\sqrt{p}\,|G_p(\alpha) - \bar{G}_p(\alpha)| \geq \lambda\} \\ &\leq 2\exp\left\{-\frac{\lambda^2}{2\sigma^2}\psi\left(\frac{2\lambda}{\sigma^2\sqrt{p}}\right)\right\}, \end{aligned} \tag{A.35}$$

where $\psi(\lambda) = (2/\lambda^2)\{(1+\lambda)\log(1+\lambda)-1\}$ is monotonely decreasing in λ and satisfies $\lambda^2\psi(\lambda) \sim 2\lambda\log\lambda$ for large λ, and σ^2 is the average variance of X_j:

$$\sigma^2 = \frac{1}{p}\sum_{j=1}^{p}\Big[P(T^{(j)}_{c^{(j)}} > \hat{t}^{(j)}_\alpha) - \big\{P\big(T^{(j)}_{c^{(j)}} > \hat{t}^{(j)}_\alpha\big)\big\}^2\Big].$$

On one hand, recall that there is at least a fraction $(1-\epsilon_n)$ of $c^{(j)}$s that are 0, and that when $c^{(j)} = 0$, $P(T^{(j)}_{c^{(j)}} > \hat{t}^{(j)}_\alpha) \sim \alpha$. We see that

$$\sqrt{p}\,\sigma \gtrsim \sqrt{p}\,\sqrt{\alpha} \geq 1. \tag{A.36}$$

On the other hand, by Schwartz inequality,

$$\sigma^2 \leq \frac{1}{p}\sum_{j=1}^{p} P\big(T^{(j)}_{c^{(j)}} > \hat{t}^{(j)}_\alpha\big) - \Big\{\frac{1}{p}\sum_{j=1}^{p} P\big(T^{(j)}_{c^{(j)}} > \hat{t}^{(j)}_\alpha\big)\Big\}^2 = \bar{G}_p(\alpha)\{1-\bar{G}_p(\alpha)\}.$$

It follows from the definition of λ that

$$\frac{\lambda}{\sigma} \geq C\log p. \tag{A.37}$$

Recalling that ψ is monotonely decreasing, and that $\lambda^2\psi(\lambda) \sim 2\lambda\log(\lambda)$ for large λ, it follows from (A.36)–(A.37) that

$$2\text{exp}\Big\{-\frac{\lambda^2}{2\sigma^2}\psi\Big(\frac{2\lambda}{\sigma^2\sqrt{p}}\Big)\Big\} \leq 2\text{exp}\Big\{-\frac{\lambda^2}{2\sigma^2}\psi\Big(\frac{C\lambda}{\sigma}\Big)\Big\} \leq 2\text{exp}\Big[-C\Big\{\frac{\lambda}{\sigma}\log\Big(\frac{\lambda}{\sigma}\Big)\Big\}\Big], \tag{A.38}$$

where $C > 0$ is a generic constant. Note that the last term in (A.38) $= o(1/p)$. Combining (A.34)–(A.35) and (A.37)–(A.38) gives (A.32).

It remains to prove (A.33). Recall that $\bar{G}_p(\alpha) = o(1)$. By the definition of II,

$$II = \max_{\alpha=i/p, 1\leq i\leq\alpha_0^* p}\sqrt{\frac{\bar{G}_p(\alpha)(1-\bar{G}_p(\alpha))}{\alpha(1-\alpha)}} \lesssim \max_{\alpha=i/p, 1\leq i\leq\alpha_0^* p}\sqrt{\frac{\bar{G}_p(\alpha)}{\alpha}}. \tag{A.39}$$

Under the null, by Theorem 1, $\bar{G}_p(\alpha)/\alpha = 1 + O(\sqrt{\log p}/\sqrt{n}) + o(1)$, which gives the first assertion in (A.33). For the second assertion, write

$$\frac{\bar{G}_p(\alpha)}{\alpha} = 1 + \frac{\bar{G}_p(\alpha)-\alpha}{\alpha} = 1 + \widetilde{\text{hc}}_{n,\alpha}\frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{p}\,\alpha}\ .$$

Noting that $\sqrt{\alpha(1-\alpha)}/(\sqrt{p}\,\alpha) \leq 1$, it follows that

$$\frac{\bar{G}_p(\alpha)}{\alpha} \leq 1 + |\widetilde{\text{hc}}_{n,\alpha}|. \tag{A.40}$$

Combining (A.39) and (A.40) gives the second claim of (A.33).

Consider **(B)**. In this case, the null hypothesis is true and all $c^{(j)}$s equal 0. By the definition and (A.29)–(A.31),

$$|\widetilde{\mathrm{hc}}_{n,\alpha}| \leq \frac{\sqrt{p}\,|\bar{G}_p(\alpha) - \alpha|}{\sqrt{\alpha(1-\alpha)}} \lesssim C\sqrt{\alpha \cdot \log p \cdot (p/n)} + o(1). \tag{A.41}$$

Recalling that $\alpha \leq \alpha_0^*$ with $\alpha_0^* = (n/p)\log p$, $\widetilde{\mathrm{hc}}_{n,\alpha} \leq C \log p$ and the claim follows.

Consider **(C)**. In this case, the alternative hypothesis is true, and a fraction $(1-\epsilon_n)$ of $c^{(j)}$s is 0, with the remaining of them equal to τ_n. Using (2.13), where $c = \tau_n$,

$$P(T_{c^{(j)}} > \hat{t}_\alpha^{(j)}) = (1 + o(1)) \cdot P(T_{c^{(j)}} > t_\alpha) + o(1/p) = \bar{\Phi}(z_\alpha - \tau_n) + o(1/p), \tag{A.42}$$

where $\bar{\Phi} = 1 - \Phi$ is the survival function of a $\mathrm{N}(0,1)$. Combining (A.29) and (A.42),

$$\bar{G}_p(\alpha) = (1-\epsilon_n)\alpha(1 + O(\sqrt{\log p}/\sqrt{n}) + \epsilon_n L_p \bar{\Phi}(z_\alpha - \tau_n) + o(1/p),$$

and it follows from direct calculations that

$$\begin{aligned}
\widetilde{\mathrm{hc}}_{n,\alpha} =& \frac{\sqrt{p}\,[\bar{G}_p(\alpha) - \alpha]}{\sqrt{\alpha(1-\alpha)}} \\
=& \frac{\sqrt{p}[\epsilon_n L_p \bar{\Phi}(z_\alpha - \tau_n) + (1-\epsilon_n)\alpha(1 + O(\sqrt{\log(p)/n})) - \alpha + o(1/p)]}{\sqrt{\alpha(1-\alpha)}} \\
=& \frac{L_p\sqrt{p}\,\epsilon_n \bar{\Phi}(z_\alpha - \tau_n)}{\sqrt{\alpha(1-\alpha)}} - \sqrt{p}\epsilon_n\sqrt{\frac{\alpha}{(1-\alpha)}} + O(\sqrt{p \log p\, \alpha/n}) + o(1). \qquad \text{(A.43)}
\end{aligned}$$

Recall that $\alpha \leq \alpha_n^* = n\,p^{-1}\log p$. First, $\sqrt{p}\epsilon_n\sqrt{\alpha/(1-\alpha)} \lesssim \epsilon_n\sqrt{p\alpha_0^*} \leq \epsilon_n\sqrt{n\log(p)}$. This equals $L_p p^{\theta/2-\beta} = o(1)$ because $\theta < 1$ and $\beta > \frac{1}{2}$. Second, $\sqrt{p\log p \cdot \alpha/n} \leq \sqrt{p\log(p)\alpha_0^*} \leq \log p$. Inserting these into (A.43) gives

$$\widetilde{\mathrm{hc}}_{n,\alpha} = \frac{L_p\sqrt{p}\,\epsilon_n\bar{\Phi}(z_\alpha - \tau_n)}{\sqrt{\alpha(1-\alpha)}} + L_p,$$

and so

$$\widetilde{\mathrm{hc}}_n^* = III + L_p, \qquad \text{where } III = \max_{\alpha = i/p, 1 \leq i \leq \alpha_0^* p} \frac{L_p\sqrt{p}\,\epsilon_n\bar{\Phi}(z_\alpha - \tau_n)}{\sqrt{\alpha(1-\alpha)}}. \tag{A.44}$$

We now re-parametrize with z_α as

$$z_\alpha = \sqrt{2q\log p} \equiv s_n(q), \text{ where } q > 0, \text{ so that } \qquad \alpha = \bar{\Phi}\{s_n(q)\}.$$

By Mill's ratio, we have $\bar{\Phi}(s_n(q)) = L_p p^{-q}$. Recall that $1/p \leq \alpha \leq \alpha_0^*$, where $\alpha_0^* = L_p p^{\theta-1}$. We deduce that the range of possible values for the parameter q runs from $(1-\theta)$ to 1 (with lower order terms neglected). It follows from elementary calculus that

$$III = \max_{(1-\theta)\leq q\leq 1} \frac{L_p\sqrt{p}\,\epsilon_n\bar{\Phi}\{s_n(q)-\tau_n\}}{\sqrt{\bar{\Phi}\{s_n(q)\}[1-\bar{\Phi}\{s_n(q)\}]}} = L_p \cdot \max_{(1-\theta)\leq q\leq 1} \frac{\sqrt{p}\,\epsilon_n\bar{\Phi}\{s_n(q)-\tau_n\}}{p^{-q/2}}. \tag{A.45}$$

Moreover, by Mill's ratio,

$$\sqrt{p}\,\epsilon_n\bar{\Phi}\{s_n(q)-\tau_n\} = L_p \cdot p^{\pi(q,\beta,r)}, \tag{A.46}$$

where

$$\pi(q,\beta,r) = \begin{cases} \frac{1}{2}-\beta, & 0<q<r, \\ \frac{1}{2}-\beta-(\sqrt{q}-\sqrt{r})^2, & r<q<1. \end{cases}$$

Inserting (A.46) into (A.45) gives

$$III = L_p \cdot \max_{(1-\theta)\leq q\leq 1} p^{\pi(q;\beta,r)+q/2}. \tag{A.47}$$

We now analyze $\pi(q;\beta,r)+q/2$ as a function of $q\in(0,1]$. In region (i), $4r\leq(1-\theta)$, and $\pi(q;\beta,r)+q/2$ is monotonely decreasing in $[(1-\theta),1]$. Therefore, the maximizing value of q is $(1-\theta)$, at which $\pi(q;\beta,r)+q/2 = \frac{1}{2}-\beta+(1-\theta)/2-\{\sqrt{(1-\theta)}-\sqrt{r}\}^2$. In region (ii), $(1-\theta)<4r\leq 1$. As q ranges between $(1-\theta)$ and 1, $\pi(q;\beta,r)+q/2$ first monotonely increases and reaches the maximum at $q=4r$, then monotonely decreases. The maximum of $\pi(q;\beta,r)+q/2$ is then $r-\beta+\frac{1}{2}$. In region (iii), $4r>1$, and $\pi(q;\beta,r)+q/2$ is monotonely increasing in $[(1-\theta),1]$. The maximizing value of q is 1, at which $\pi(q;\beta,r)+q/2 = 1-\beta-(1-\sqrt{r})^2$. Combining these with (A.47) and (A.44) gives the claim. □

References

SHORACK, G.R. AND WELLNER, J.A. (1986). Empirical Process with Application to Statistics. *John Wiley & Sons, NY.*